Domain Name Valuation: Characteristics & Price Exposed!

A thesis submitted to the
Graduate School of Natural and Applied Sciences

by

Emrullah DELİBAŞ

in partial fulfillment for the
degree of Master of Science

in
Data Science

İSTANBUL
ŞEHİR
UNIVERSITY

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Data Science.

APPROVED BY:

Asst. Prof. Barış Arslan ....
(Thesis Advisor)

Asst. Prof. Ali Çakmak ....

Asst. Prof. Tevfik Aytekin ....

This is to confirm that this thesis complies with all the standards set by the Graduate School of Natural and Applied Sciences of İstanbul Şehir University:

DATE OF APPROVAL: 29.08.2019

SEAL/SIGNATURE:

Declaration of Authorship

I, Emrullah DELİBAŞ, declare that this thesis titled, 'Domain Name Valuation: Characteristics & Price Exposed!' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: 29 August 2019

For indeed, with hardship [will be] ease.

Qur'an 94:6

Domain Name Valuation:
Characteristics & Price Exposed!

Emrullah DELİBAŞ

Abstract

Given only the domain name, can we predict its price? This is the main question that is examined within the scope of this thesis. Price prediction is one of the very well studied applications of machine learning (ML). An accurate ML approach for price prediction would need a good set of features to represent characteristics that effect the price. Price of a domain name depends not only on its characteristics such as length, language and extension, but also how much a person is willing to pay for it. This introduces a significant uncertainty in domain name valuation and creates a challenging problem to deal with. Additionally, domain names are in a special form of an unseparated text that can only consist of letters, numbers, hyphens and emojis, and with an additional structural limitation on having length of sixty-three characters at most in its puny-coded representation. Exposing all decisive characteristics of a domain name that affect its value consequently proves to be a very challenging task. An extensive domain name sale history dataset is collected as part of this study and numerous unique features are extracted based on domain name.

One of the crucial steps in feature extraction is the identification of words in the domain name. This process includes language identification and word segmentation step that is referred to as Domain Name Language Detection (DNLD) in this thesis work. Identification of domain language and the extraction of words within a domain name is essential in representing the domain name characteristics that have profound effect on its value such as the number of words and the popularity of words used in domain name. DNLD utilizes Fasttext dataset from Facebook [1, 2] and can support up to 265 languages.

Keywords: Domain Name, Domain Appraisal, Domain Valuation, Language Identification, Word Segmentation, Price Estimation, Machine Learning, Natural Language Processing

Alan Adı Değer Tespiti:
Özellikler ve Fiyatın Açığa Çıkarılması!

Emrullah DELİBAŞ

Öz

Yalnızca alan adı göz önüne alındığında, o alan adının fiyatını tahmin edebilir miyiz? Bu tez kapsamında ele alınan asıl soru budur. Fiyat tahmini, makine öğrenmesinin (ML) çok iyi çalışılmış uygulamalarından biridir. Fiyat tahmini için doğru bir ML yaklaşımı, fiyatı etkileyen özellikleri yansıtmak için iyi bir dizi özelliğe ihtiyaç duyacaktır. Bir alan adının fiyatı sadece uzunluk, dil ve uzantı gibi özelliklerine değil, aynı zamanda kişinin o alan adı için ne kadar ödemeyi göze aldığına da bağlıdır. Bu, alan adı değerlemesinde ciddi bir belirsizliğe neden olur ve başa çıkılması zor bir sorun yaratır. Ek olarak, alan adları yalnızca harfleri, sayıları, tire ve emojileri içerebilen, bitişik yazılan ve punycode gösteriminde en fazla altmış üç karakter uzunluğunda olabilen yapısal bir sınırlamaya sahiptir. Bir alan adının değerini etkileyen tüm belirleyici özelliklerinin ortaya çıkarılması netice itibariyle oldukça zor bir iştir. Bu çalışmanın bir parçası olarak kapsamlı bir alan adı satış geçmişi veri kümesi toplanmış ve alan adına bağlı çok sayıda özgün özellik elde edilmiştir.

Özellik çıkarımında en kritik adımlardan biri alan adındaki kelimelerin belirlenmesidir. Bu süreç, bu tez çalışmasında Alan Adı Dil Tespiti (DNLD) olarak adlandırılan dil tanımlama ve kelimelere bölme adımını içermektedir. Alan adı dilinin ve içindeki kelimelerin tespit edilmesi, kelime sayısı ve alan adında kullanılan kelimelerin popülerliği gibi alan adının değeri üzerinde önemli bir etkiye sahip olan özellikleri ifade etmede elzemdir. DNLD, Facebook [1, 2] tarafından paylaşılan Fasttext veri kümesini kullanır ve 265 dile kadar destekleyebilir.

Anahtar Sözcükler: Alan Adı, Alan Adı Ekspertizi, Alan Adı Değer Tespiti, Dil Belirleme, Kelimelere Bölme, Fiyat Tahmini, Makine Öğrenmesi, Doğal Dil İşleme

I would like to dedicate this work to my beloved wife, who constantly looked for opportunities to create a suitable environment for me to complete this study.

Acknowledgments

I would like to express my sincere gratitude to my advisor Asst. Prof. Barış Arslan for the continuous support during my study, and providing me an opportunity to use the laboratory. Without his patience, motivation, and immense knowledge, it would not be possible to conduct this study.

I would like to thank the rest of the committee, Asst. Prof. Ali Çakmak and Asst. Prof. Tevfik Aytekin, for their insightful comments and encouragements.

I am also grateful to Asst. Prof. Hatice Tekiner Moğulkoç, the director of the Graduate School of Natural and Applied Sciences of İstanbul Şehir University, for all her efforts to solve my problems.

I also would like to thank my colleagues, M. Tuna and Muhammet Macit, for their valuable comments.

Last but not the least, I am grateful to my family for providing me continuous encouragement throughout my master studies.

Contents

List of Figures

List of Tables

Abbreviations

API	**A**pplication **P**rogramming **I**interface
ccTLD	**c**ountry **c**ode **T**op-**L**evel **D**omain
CV	**C**ross **V**alidation
DNLD	**D**omain **N**ame **L**anguage **D**etection
DNPE	**D**omain **N**ame **P**rice **E**stimation
DNS	**D**omain **N**ame **S**ervice
gTLD	**g**eneric **T**op-**L**evel **D**omain
IDN	**I**nternationalized **D**omain **N**ame
IP	**I**nternet **P**rotocol
IQR	**I**nter**q**uartile **R**ange Method
LI	**L**anguage **I**dentification
ML	**M**achine **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
newgTLD	**new** **g**eneric **T**op-**L**evel **D**omain
NLP	**N**atural **L**anguage **P**rocessing
STD	**St**andard **D**eviation
SVR	Epsilon-**S**upport **V**ector **R**egression
W3Techs	**W**orld **W**ide **W**eb **Tech**nology **S**urveys
XGB	E**x**reme **G**radient **B**oosting

Chapter 1

Introduction

A name, with simplistic terms, is a label that is used to separate one thing from another. For instance, human beings have names, which consist of letters inherited from a language, that allows an individual to be addressed. Likewise, computers distributed all over the world should be named in order to separate one machine from another and to enable activities such as network communication among them. To put it more precisely, in today's world, the *name* a computer needs is an Internet Protocol (IP) Address.

Internet Protocol (IP) Address – Formally can be defined as a "label assigned to each device connected to a computer network that uses the Internet Protocol for communication" [3].

Even-though they are perfectly sufficient for network based communications to occur, it is not desirable to use them directly since they are not easy to communicate for humans. Because, it would be nearly impossible to remember the IP Addresses of all your favorite web sites for instance. So, the concept of domain name came to life in order to alleviate this issue.

Domain Name – A domain name can be identified as a special form of an unseparated text that can only consist of letters, numbers, hyphens and emojis, and with an additional structural limitation of having a length of sixty-three characters at most in its punycode representation. Figure 1.1 illustrates the structure of a domain name.

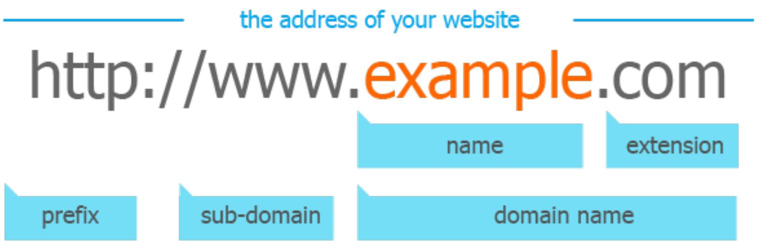


FIGURE 1.1: Structure of a domain name [4]

It can be represented as unicode as well and these two representations differ only for internationalized domain names (IDNs). This special type of usage enables people register domain names in languages other than English as defined in [5]. Table 1.1 can be examined with this in mind.

TABLE 1.1: Domain Name Representation Examples

Language	**Punycode**	**Unicode**
English	istanbulsehiruniversity.com	istanbulsehiruniversity.com
Turkish	xn--istanbulehirniversitesi-npc48v.com	istanbulşehirüniversitesi.com

Internationalized Domain Name (IDN) – Starting with the prefix "xn--" is the ultimate indication of a domain name being an IDN in its punycode representation as explained in [6–8].

The connection between an IP Address and a domain name can be understood like a mapping where domain name points to IP Address (such as sehir.edu.tr -> 91.93.32.250). In other words, domain name, in fact, functions as a symbolic link to the actual IP Address. So, in this architecture, domain name is a nickname that does contain no information other than pointing to the place where the actual information resides (i.e. IP Address). In practice, the required information to generate this mapping functionality is first kept in a file named HOST.TXT and was placed at each computer (i.e. physical address) that wishes to connect to the Internet. However, it is clear that when all computers in the world are considered, this architecture is not scalable. To overcome this issue and fulfill the need of scalability, a hierarchical database of domain names was established and named as Domain Name System (DNS), which also includes location and institution related information [9]. DNS can be considered as the phone-book of the Internet.

Domain names are much more than just a technical shortcut. According to The Domain Name Industry Brief Report issued by Verisign [10] there are approximately 351.8 million registered domain names in the world of the Internet with more than 1,500 different extensions (.com, .net, .org, etc.). With the high circulation and transaction, more than 250,000 domain names are being registered and almost 200,000 are being deleted daily. Many sectors, such as domain name registration, sales, lease, parking and other services (whois, appraisal, consultancy, etc.) are operating in the industry. Only the registry & registrar business being in the range of $3 to $5G (billion) according to recent reviews [11] provides an indication on how big the domain name industry can be.

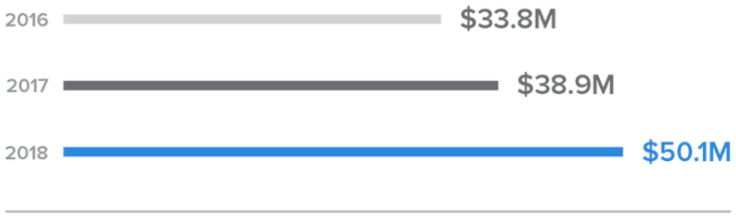


Figure 1.2: Total Sales by Year at Uniregistry [12]

There are many domain marketplaces exist such as GoDaddy, Uniregistry, Sedo, etc. Figure 1.2 and Figure 1.3 illustrate total sales and domains sold, respectively, by year (2016-2018) at Uniregistry only. Even-though these statistics are taken from a single marketplace only, it is clear that there is a continuous growth trend in domain name industry.

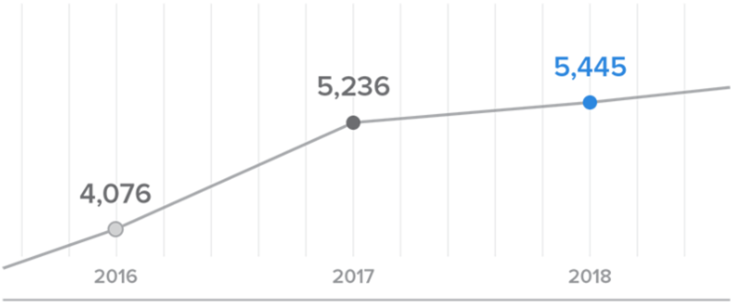


Figure 1.3: Domains Sold By Year at Uniregistry [12]

Value of a domain name depends on not only its technical properties such as length, language and extension, but also how much a person is willing to pay for it. This introduces a significant uncertainty in domain name valuation and creates a challenging problem to deal with. Considering the size of the industry, it is however quite valuable to accurately appraise a given domain name since it is vital for 1) marketplaces to come up with accurate and competitive pricing, 2) informed decision-making on personal and business domain name purchases, 3) domain name investment decisions. This can be considered as the main motivation behind the study. So, within the scope of this thesis, the question of whether it is possible to accurately valuate a domain name is examined by utilizing a Machine Learning (ML) based approach. Sub-studies such as language identification (LI), word segmentation, etc. are also performed during the keyword extraction phase of this thesis. They are very essential in terms of understanding characteristics of a domain name and crucial to the completeness and comprehensiveness of the study as well.

Other motivations behind this study are as follows:

- Classical language identification approaches known in NLP are not directly applicable since they mainly expect to be fed with respectively long and word level separated texts as input because domain names have the length of 63 at most and represented as unseparated text.
- Known LI studies mostly have support only for major languages (English, Spanish, German, French, Italian, etc.) only.
- To the best of our knowledge, there is no previously published study on domain name language identification and keyword extraction.
- Although there is a considerable number of commercial tools such as GoValue [13], Estibot [14], Epik [15] and others [16, 17] for domain price prediction, the number of previously published study on domain price prediction is quite limited.

Major contributions of this study are as follows:

- To the best of our knowledge, this is the first study that proposes a language identification and keyword extraction approach specific to domain names (DNLD).
- DNLD support up to 265 languages depending on Fasttext datasets.
- DNLD is modeled as a weighted interval scheduling problem and a number of weighting schemes, which considers word length and word's usage frequency (i.e. word popularity), are proposed.
- An extensive data collection and feature extraction effort to represent domain sale data and numerous domain characteristics.

- A ML-based price estimation model for domain names (DNPE) by utilizing DNLD outputs and numerous other domain characteristics.

The rest of this thesis is organized as follows: Chapter 2 discusses the related studies from the literature in a comprehensive manner. Chapter 3 presents overall flow. Chapter 4 explains data exploration and preparation phase. Chapter 5 introduces the methodology & modeling that we propose. Chapter 6 discusses the results of our study. Finally, Chapter 7 provides brief conclusion and Chapter 8 discusses possible future studies.

Chapter 2

Related Work

2.1 Domain Name Language Detection (DNLD)

To the best of our knowledge, there is no previously published work on identifying the language of domain names. A few studies classifies URLs based on their languages[18–20] and categories [21, 22]. There are some studies classifying URLs based on content as well [23–25]. However, a URL may contain protocol, folder, sub-folder, page, parameter, etc. related information on top of containing domain name. So, although their starting points are the same in nature, they considerably differ in terms of the actual problem they deal with, the data-sets they use, and the algorithms they apply.

The very first thing that comes to mind is the usage of stop words as language indicators [26]. The frequent occurrence of the word "the" in a text is an indication of it being English for instance. There are also derivatives of this approach considering discriminative letter sequences such as "ery_" and "eux_" for being indicators of English and French, respectively. The problem with these approaches is that they work well given a fairly long sample of text, however, they are not applicable to our setting of domain name language detection where the input is very short, unseparated and does not even have to contain any proper words.

Nigam et. al. [27] proposes a maximum entropy-based approach where they train a probabilistic model for each language based on word occurrences in documents by using an improved iterative scaling scheme. Then, language assignment of a new document is actualized to the language with highest conditional probability based on pre-calculated distribution of words in the document. We conceptually adopted a similar approach. However, word popularity distribution is modeled based on work ranking information, given no word occurrence information in Fasttext dataset and a weighted interval scheduling algorithm is used during language assignment phase along with the proposed weighting schemes.

The approaches mentioned so far mainly assume that the text given as input is long enough, properly structured and no misspelling is involved. On the other hand, domain names are short, unseparated and very convenient for misspellings. So, use of n-grams are preferred over words in these situations since it allows only the misspelled part to be affected rather than the whole word in theory. The work of Grefenstette. et. al. [28] shows that n-gram based approaches (tri-grams, specifically) outperform stop word based approaches on short texts. Later on, character-based Markov models [29] are proposed as a derivative to n-gram approach, where it is assumed that the next character depends on a certain number of previous characters only and the probability distribution of sequence of characters are generated based on this. Markov models mentioned above are also used in the work of Teahan et. al. [30] for the training phase of Prediction by Partial Match compression models, where a new document is classified by its compression performance. Later on, a graph-based n-gram model called LIGA [31] which tries to learn grammar elements by considering the order of n-grams on top of n-gram frequencies is proposed. According to results shared, LIGA outperforms classical n-gram approach. All the approaches stated so far, in one way or another, classify the document to one of the languages involved. However, document may contain keywords from multiple languages or may even be a combination of unmeaningful letters only. Řehřek et. al. [32] investigates the limitations of classical n-gram approaches, addresses the issues mentioned above and proposes a new method which constructs language models based on word relevance. We adopted the use of ngrams in the generation phase of possible intervals, which will be provided as input to weighted interval scheduling algorithm so that the language is decided among the language candidates offering the best schedule possible (i.e. set of non-overlapping intervals).

One thing in common for all the n-gram based approaches is that they all start with building n-gram distribution for the languages involved in classification task and end with assigning the document to the language having the "most similar" n-gram distribution with exception that the number of n-grams to be used and how the "similarity" is defined.

It is possible to choose *all* trigrams directly since it outperforms the others as suggested in [20] but either the k most frequent n-grams or all n-grams which occur more than k times are candidate to be chosen as suggested in [33]. Hayati et. al. [34] discusses the issue of how the selection of n-grams should be done in detail. In terms of how the "similarity" is defined, Cavnar et. al. [33] proposes a aforementioned rank-order statistic model which compares the different frequency ranks. Using Relative Entropy as a distance measure is also proposed in [35]. Several distance measures are experimented and performance comparison results are reported in [18, 19, 23]. The problem with most of the approaches mentioned so far is that they are mainly experimented with major languages (English, Spanish, German, French, Italian, etc.) only.

Word segmentation is not as much studied as language identification. It is also more challenging compared to LI in terms of complexity. We are not aware of any prior work on domain name keyword extraction in the literature. However, since a domain name is an unseparated text in nature, [36] is worth mentioning as it provides various probabilistic models trained according to the frequency of the usage of words in the language to decide on best combination of keywords and extract them with additional spelling correction approaches. It does not however satisfy our requirements as the algorithmic approaches are quite simplistic, and it only works for one language and is designed for longer texts compared to a domain name, which is limited with 63 characters only.

2.2 Domain Name Price Estimation (DNPE)

To the best of our knowledge, there is only a limited number of studies on domain name appraisal [37–39]. The scope of Bikadi et. al.'s work is limited to classifying whether a domain name is valuable using Support Vector Machines for 903 domain names [37]. A few methods based on case-based reasoning and artificial neural networks are proposed, where the evaluation is done only with 4,231 domain name transactions in comparison to 587,377 domain names in our study and the best approach is resulted with an R^2 score of 0.532 [38, 39]. A number of previously published work [9, 40–52] studies the factors that potentially effect the value of domain names but they do not develop models to predict the domain name price.

The value of a domain name is highly effected by its structural properties. Domain names with gTLD extensions (com, net, org, etc.) are preferred over ccTLD extensions (tr, de, fr, etc.) most of the time. So, generally a gTLD is higher in value than ccTLD [10, 44, 53]. Shorter domains are generally more valuable than the others since they are easier to memorize, recall, spell, etc. [9, 45]. Words, potentially having meaning in a multi-language manner are very crucial in terms of their commercial significance [46, 47]. Mueller et. al supports this argument by revealing how domain names can effect marketing with search engine results [48]. Statistically, domain name itself is the one influences the user to click the URLs appeared in search results for about one in four cases [49]. Many of these factors that effect the domain value are used as features in our domain name prediction model and, as can be understood from these studies, the keywords within a domain name (and potentially its language) are quite decisive for its valuation. So, DNLD was developed as a natural result of this need.

Brand relevance of a domain name is another factor may have impact on its value. Consider name of a company being the same as its domain name for instance, this will most likely boost the value of its domain name since it makes customer's job easier [50, 51]. Rarity also has effect on domain name's value. For example, some of one character domains such as q.com have been frozen by ICANN, which caused their value to increase [41].

Culture, beliefs and social trends can also influence the value of the domain name [41]. A domain name may gain different meanings from culture to culture. For example, the numbers 168 and 1314 have special meanings Chinese; "good luck all the way" and "a lifelong love" respectively. So their value will be higher compared to other numbers registered. A domain name that refers to a new technology accepted by society may also be more valuable than its peers for at least a certain period of time.

All the works mentioned above provide properties that can affect the value of the domain name in one way or another. In this study, we did our best to characterize these properties and transform them into features as much as possible. Those who could not be represented as features at the moment are recommended as future work.

Chapter 3

Overall Flow

In this chapter, we review each step involved in domain evaluation process as depicted in Figure 3.1.

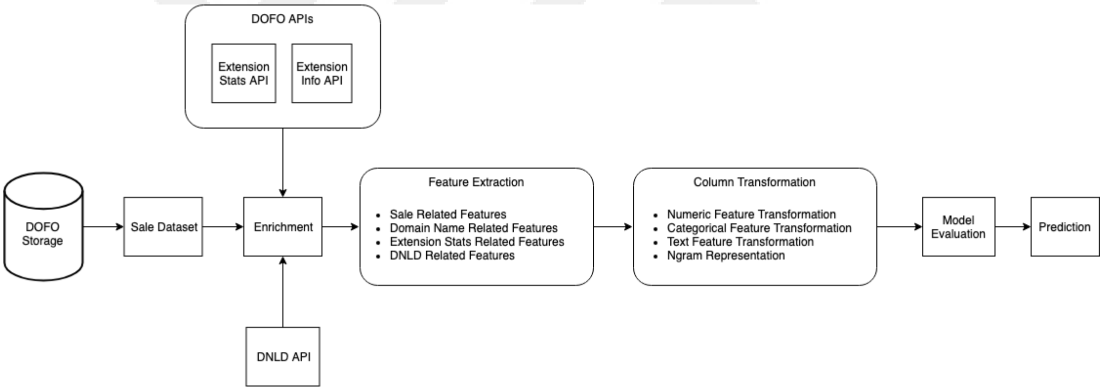


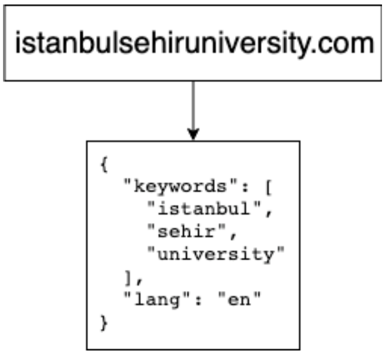
Figure 3.1: Domain Evaluation Flow

The primary data available in the domain name price estimation is historical data of domain sales. In this study, the required sale dataset is enabled and exported from DOFO's [54] Storage, which is constantly updated as a natural consequence of a unique and well designed data processing pipeline construction. Details of the pipeline, dataset preparation and feature extraction in particular, are explained in Chapter 4. Diversity of the data and a large set of features are essential in machine learning tasks in order to uncover the patterns in data. Furthermore, existence of a diverse dataset paves the way for the generation of new features that helps with machine learning model training. Consequently, enrichment processes on exported sale dataset is applied in three steps:

1. We enriched the data with meta information such as domain length, IDN status, extension type, etc. extracted for both name and extension sides of provided domain names.
2. We collected statistical information such as extension popularity, registered extension count, etc. for both name and extension sides of provided domain names from DOFO APIs.
3. We enriched with language and keyword related information as a result of language identification and word segmentation sub-studies that are completed within the scope of this study.

Language identification (LI) is classified as a Natural Language Processing (NLP) task and it is one of the well studied topics within the field. Although there is a number of study done on it already, most of these studies start with the assumption of having relatively long text as input. Since the scope of this study is based on domains, which are limited with 63 characters at most and represented as unseparated text, the approaches that they adopted cannot be directly applicable in this case. Moreover, DNLD is very important in terms of understanding characteristics of a domain name since it enables us to identify its language, extracts keywords within and derive various new features on top of them, so it is essential to the completeness of the study. So, we modeled it as a weighted interval scheduling problem and come up with a solution that is based on dynamic programming. More specifically, we propose a model to compute 'weight' and use dynamic programming to schedule weighted intervals that are generated with meaningful ngrams of provided domain name. Weight in our implementation depends on word length, word rank and word count by language. Thus, our model has the capability to decide on the language and extract the keywords within a domain name. This model is trained with Fasttext datasets, which has a language scale of 265, from Facebook [1, 2]. Figure 3.2 shows an example output of DNLD. Additional details on DNLD will be explained in Chapter 5. And evaluation results will be presented in Chapter 6.

Enrichment process is then continued with feature extraction and selection. This step can be considered as the center of the study because no good prediction come to life without highly qualified and selective features provided. So, with the addition of DNLD enrichment, features extracted in this step can be categorized into four groups: 1) sale related features such as venue (i.e. marketplace), sale weekday, sale season, etc. 2) domain name related features such as domain length, letter count, extension, extension



```
{
  "keywords": [
    "istanbul",
    "sehir",
    "university"
  ],
  "lang": "en"
}
```

FIGURE 3.2: An example to DNLD

type, extension popularity, etc. 3) extension stats related features such as registered extension count, etc. 4) DNLD related features such as domain language, keywords, number of keywords, avg. keyword popularity, etc.

Most of machine learning regression models and implementations do not work with non-numeric inputs, so they mainly expect to be fed with floating inputs. Because of this requirement, all features extracted in the previous step categorized based on their types first and then transformed accordingly. Applied column transformations can be divided into four groups:

1. **Numeric feature transformation**: applied on ready-to-go numeric features such as domain length, avg. keyword popularity, number of keywords, etc. for scaling purposes only.
2. **Categorical feature transformation**: applied on text features that can be considered as categorical and not having any form of hierarchy within such as venue (i.e. marketplace), domain language, IDN status, etc.
3. **Text feature transformation**: applied on text features that are valuable by itself such as domain keywords.
4. **Ngram representation**: applied on text features that are valuable when it is divided into its ngrams such as name side of a domain name.

Pre-processed data is then split as train (80%) and test (20%) sets and 3-fold cross validation (CV) is utilized for hyper-parameter tuning and model selection. Most of known regression models such as Linear Regression, Support Vector Regression (SVR), Multi-Layer Perceptron Regression (MLPRegressor), Random Forest Regressor and Extreme Gradient Boosting Regression (XGBRegressor) are evaluated during the process.

Data exploration and preparation related operations such as collection, wrangling and feature extraction are explained in Chapter 4. Details of DNLD and price estimation methodology & modeling is presented in Chapter 5 and results are discussed in Chapter 6.

Chapter 4

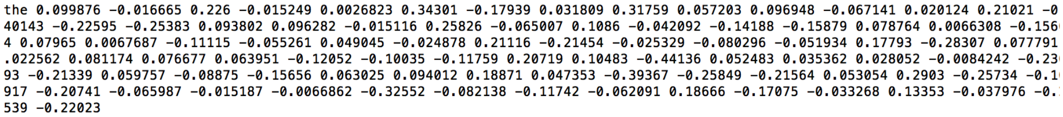
Data Preparation

4.1 Data Collection

4.1.1 DNLD Datasets

4.1.1.1 Fasttext Datasets

Fasttext datasets consist of pre-trained word vectors (i.e. words and their vectorial representations), trained in 300 dimensions for multiple languages. A sample output of Fasttext vector for the word *the* is as illustrated in Figure 4.1.



```
the 0.099876 -0.016665 0.226 -0.015249 0.0026823 0.34301 -0.17939 0.031809 0.31759 0.057203 0.096948 -0.067141 0.020124 0.21021 -0
40143 -0.22595 -0.25383 0.093802 0.096282 -0.015116 0.25826 -0.065007 0.1086 -0.042092 -0.14188 -0.15879 0.078764 0.0066308 -0.156
4 0.07965 0.0067687 -0.11115 -0.055261 0.049045 -0.024878 0.21116 -0.21454 -0.025329 -0.080296 -0.051934 0.17793 -0.28307 0.077791
.022562 0.081174 0.076677 0.063951 -0.12052 -0.10035 -0.11759 0.20719 0.10483 -0.44136 0.052483 0.035362 0.028052 -0.0084242 -0.23
93 -0.21339 0.059757 -0.08875 -0.15656 0.063025 0.094012 0.18871 0.047353 -0.39367 -0.25849 -0.21564 0.053054 0.2903 -0.25734 -0.1
917 -0.20741 -0.065987 -0.015187 -0.0066862 -0.32552 -0.082138 -0.11742 -0.062091 0.18666 -0.17075 -0.033268 0.13353 -0.037976 -0.
539 -0.22023
```

Figure 4.1: Sample Output of Fasttext Vector

As summarized in Table 4.1, first Fasttext model (Fasttext-v1) [1] is trained with Wikipedia only and contains 18,912,933 domain compatible distinct words for 265 languages in total, whereas second Fasttext model (Fasttext-v2) [2] is trained with Common Crawl as well and contains 81,221,971 domain compatible distinct words for 157 languages in total. It should be noted that even though the number of distinct words in the second study is a lot more, the number of languages is lower. The reason behind the elimination of languages is that the words in those languages are often the duplicates of the words in major languages and therefore do not contain enough unique words to be considered as a separate language. These datasets are publicly available and can be downloaded from the official website of Fasttext Project [55].

Table 4.1: Fasttext Dataset Information

Dataset	**Lang. Scale**	**Word Count**	**E. Word Count**	**Trained with**
Fasttext-v1	265	18,912,933	27,612,931	Wikipedia
Fasttext-v2	157	81,221,971	108,263,268	Wikipedia, C. C.

In this study, Fasttext datasets are used to provide required look-up datasets in DNLD, which will be explained Data Wrangling Section. The main reason behind choosing these datasets is the support for a exceedingly large number of languages.

4.1.1.2 Noktadomains.com Dataset

Noktadomains.com is a platform specialized on domaining and functions as domain sales service mostly. The importance of this platform in terms of this study is that it provides language and keyword support for 7 languages (We could not find any other platform providing this information). So, a total of 196,832 DNLD related domain information is crawled from [56], which will be used during evaluation step of DNLD. Supported languages and their document count statistics can be found at Table 4.2.

Table 4.2: Noktadomains.com Dataset Doc. Distribution by Language

Language	**Doc. Count**
english	190,714
turkish	5,040
german	476
spanish	369
french	171
italian	38
dutch	24

4.1.2 Sale Dataset

Sale dataset is retrieved from DOFO [54]. General architecture of DOFO's data processing pipeline is shown in Figure 4.2.

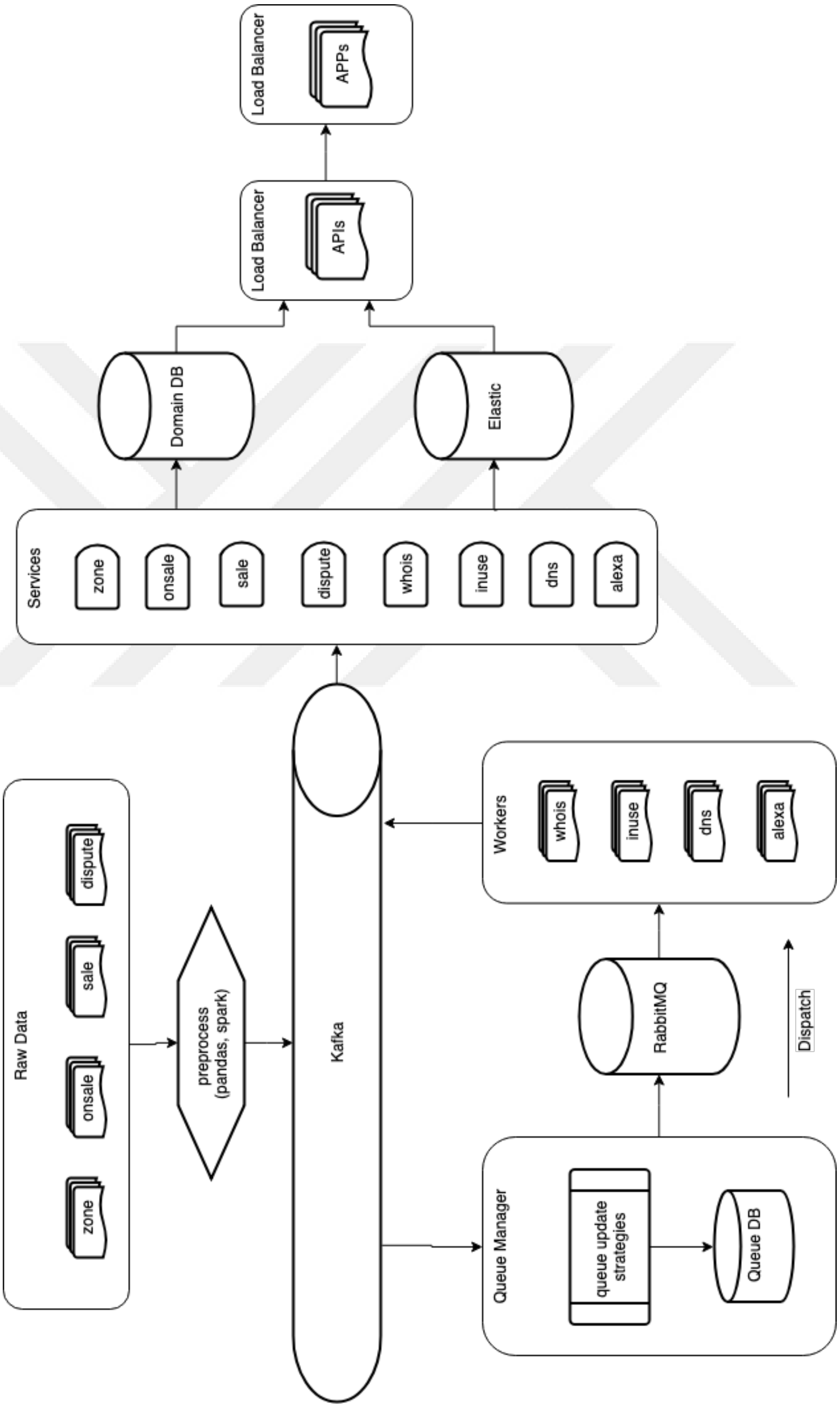


Figure 4.2: General Architecture of Data Processing Cycle at DOFO

DOFO is a platform specialized on domain names. Their motto is *google of domain names.* So, they collect zone, onsale, sale and dispute data for all domain names (~350M) from various platforms at a daily bases and apply batch processing on top of them. Additionally, they apply real time processing for whois, inuse, dns, alexa, etc. related data in order to keep their storage up-to-date. Since the amount of data is huge, they use big data technologies such as Apache Spark [57], Apache Kafka [58], etc., and apply Lambda Architecture [59] on top of that in order to speed up processing in a cost-effective manner. The structure of sale dataset is illustrated in Table 4.3.

Table 4.3: Raw Sale Dataset Sample

Domain	**Venue**	**Price**	**Date**
cowpies.com	snapnames	1,750	2005-01-28
kidsong.com	snapnames	1,750	2005-01-20
2dc.net	afternic	750	2005-02-15
fontalicious.com	snapnames	1,750	2005-12-30
sex.com	private	14,000,000	2005-01-01

Retrieved sale dataset consists of 587,377 rows. Minimum and maximum prices are 90$ and 14,000,000$ respectively with an average of 2,640.935 as stated in Table 4.4.

Table 4.4: Domain Sale Price Stats

	price
count	587,377
mean	2,640.935
std	51,224.7
min	90
25%	200
50%	500
75%	1,710
max	14,000,000

The low numbers correspond to 25%, 50% and 75% is a natural indication on how price distributes across the sales. To be more specific, the gap between min and max is dramatically huge and most of the sales can be considered as 'low' if we were to categorize them. The ones near maximum can be considered as anomalies and they can even be ignored during the evaluation. In Figure 4.3, domain prices are shown in log scales.

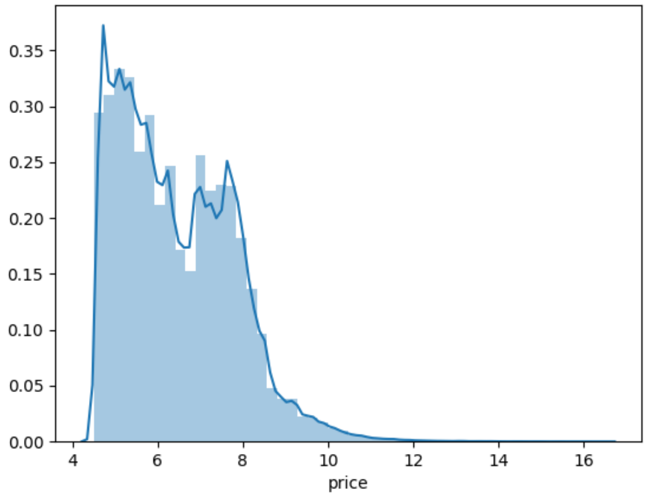


Figure 4.3: Domain Sale Price Distribution (Log)

4.2 Data Wrangling

4.2.1 Preparing DNLD Look-up Datasets

Vector based Fasttext models mentioned above consist of words and their vectorial representation in 300 dimensions. It should be noted that vector related information will not be used within the scope of this thesis, but only words. And the reason behind choosing this dataset and using it with that context is that no other dataset supporting this much language scale could be found. However, the words that is contained in this dataset may not compatible to be a part of a domain name all the time since a domain name can only consist of letters, numbers, hyphens and emojis and with the maximum length of sixty-three characters in its puny-coded form as mentioned before. So, model files are processed in order to extract the set of compatible words by converting them to puny-coded form first and then applying the appropriate regex. As stated in the documentation of Fasttext that even-thought there is no frequency information involved in the models, words are sorted by their occurrence counts. So, the language look-up datasets needed for DNLD, which will be explained in Chapter 5, are created by keeping track of word rank and word count information for each word in a by language manner as illustrated below:

$$LANG\_LOOKUP : Map < lang, word\_count >$$

$$WORD\_LOOKUP : Map < word, Map < lang, word\_rank >>$$

First is language look-up dataset that contains word count information of 265 language for Fasttext-v1 and 157 for Fasttext-v2 respectively. Second is word look-up dataset, which contains word rank information for each language that a word belongs to, of 18,912,933 words for Fasttext-v1 and 81,221,971 for Fasttext-v2 respectively.

It is quite common in domain industry that domains normally containing non-English characters are registered with a latinization process being applied first. Because this increases their appraisal and web traffic as a natural outcome. Consider a Turkish domain name, İstanbul Şehir Üniversitesi for instance, it normally should be istanbulşehirüniversitesi.com, but registering it as istanbulsehiruniversitesi.com (by converting "ş" to "s" and "ü" to "u") is actually increases its value. So, we applied this latinization process to words look-up dataset and increased total number of look-up words as a natural outcome, which will eventually expect to increase the accuracy of our DNLD approach. Thus, extended word count become 27,612,931 from 18,912,933 for Fasttext-v1 and 108,263,268 from 81,221,971 for Fasttext-v2.

4.3 Feature Extraction

The original sale dataset only contains domain name, venue, sale price and sale date related information. However, these may not be sufficient since an accurate ML approach for price prediction would need a good set of features to represent characteristics that effect the price. So, a number of features extracted and divided into 4 categories. 1) sale related features such as venue (i.e. marketplace), sale weekday, sale season, etc. 2) domain name related features such as domain length, letter count, extension, extension type, extension popularity, etc. 3) extension stats related features such as registered extension count, etc. 4) DNLD related features such as domain language, keywords, number of keywords, avg. keyword popularity, etc.

4.3.1 Sale Related Features

This indicates features that are primarily related with the sale itself such as marketplace, sale month, etc.

4.3.1.1 Venue (i.e. Marketplace)

A total of 632 distinct venue exists in the dataset. Table 4.5 provides top 5 venue. It is interesting to note that, GoDaddy is the most popular venue, but has the lowest median. We believe that median difference between marketplaces is an indicator of this feature may be useful.

Table 4.5: Top 5 Venue with Stats

Venue	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
godaddy	181,691 (0.309)	100.00	499.62	205.00	380,000.00
namejet	119,887 (0.204)	100.00	1,049.34	393.00	929,000.00
sedo	115,794 (0.197)	90.00	3,980.34	1,600.00	13,000,000.00
afternic	58,250 (0.099)	100.00	2,712.41	2,000.00	1,525,000.00
dropcatch	30,869 (0.053)	100.00	526.07	254.00	220,950.00

4.3.1.2 Sale Date

A number features are extracted based on sale date in order to reveal various hidden correlations.

4.3.1.2.1 Sale Month

Sale month indicates the month in which sale took place. According to Table 4.6, the median value of domain sales in January, which is also the most popular month, is higher than others. Additionally, median price distribution is slightly increased as we move towards the first months of the year as observed in Figure 4.4. So, these support the conclusion that month information is effective on sales.

TABLE 4.6: Sale Months with Stats

Sale Month	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
January	55,699 (0.095)	90.00	3,570.28	790.00	14,000,000.00
February	52,839 (0.090)	90.00	3,090.92	650.00	8,888,888.00
March	55,368 (0.094)	100.00	2,589.77	506.00	9,999,950.00
April	52,441 (0.089)	100.00	2,367.00	500.00	3,600,000.00
May	50,717 (0.086)	100.00	2,358.36	425.00	7,500,000.00
June	47,560 (0.081)	100.00	2,667.24	455.00	9,500,000.00
July	46,119 (0.079)	100.00	2,433.24	485.00	2,900,000.00
August	46,275 (0.079)	100.00	2,050.68	415.00	1,000,000.00
September	45,020 (0.077)	100.00	2,326.45	473.00	4,700,000.00
October	44,723 (0.076)	100.00	2,308.22	476.00	2,000,000.00
November	44,743 (0.076)	100.00	2,940.94	490.00	13,000,000.00
December	45,873 (0.078)	100.00	2,798.88	470.00	6,784,000.00

Figure 4.4 functions as the visual representation of Table 4.6.

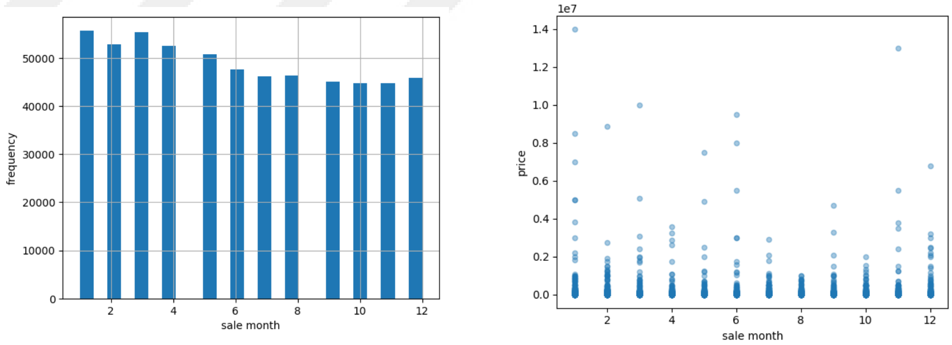


FIGURE 4.4: Histogram and Scatter Plot of Sale Month

4.3.1.2.2 Sale Season

Sale season indicates the season in which sale took place. According to Table 4.7, spring is the most popular season in terms of sales, but median value of domain sales is at its highest in winter. The fact that the seasons are effective on sales can be observed on Figure 4.5 as well.

Table 4.7: Sale Seasons with Stats

Sale Season	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
Winter	154,411 (0.263)	90.00	3,177.08	615.00	14,000,000.00
Spring	158,526 (0.270)	100.00	2,442.04	490.00	9,999,950.00
Autumn	134,486 (0.229)	100.00	2,524.82	480.00	13,000,000.00
Summer	139,954 (0.238)	100.00	2,386.27	453.00	9,500,000.00

Figure 4.5 functions as the visual representation of Table 4.7.

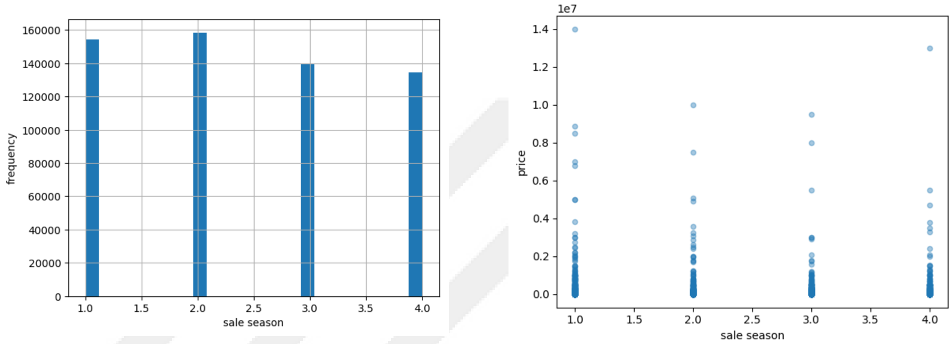


Figure 4.5: Histogram and Scatter Plot of Sale Season

4.3.1.2.3 Sale Weekday

Sale weekday indicates the day of the week in which sale took place. Table 4.8 indicates that Wednesday is the most preferred day of the week for domain sales. Median value is also at its highest in this day. And it is interesting to note that Saturday is the least preferred day. So, it is really interesting to observe that the days of the week can be so effective on domain name valuation.

Table 4.8: Sale Weekdays with Stats

Sale Weekday	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
Monday	86,944 (0.148)	100.00	1,636.42	500.00	3,300,000.00
Tuesday	97,839 (0.167)	90.00	2,915.21	623.00	9,999,950.00
Wednesday	127,601 (0.217)	100.00	4,033.82	1,200.00	13,000,000.00
Thursday	80,869 (0.138)	100.00	2,303.30	400.00	7,000,000.00
Friday	72,765 (0.124)	100.00	1,948.23	355.00	8,500,000.00
Saturday	53,974 (0.092)	90.00	2,154.02	300.00	14,000,000.00
Sunday	67,385 (0.115)	100.00	2,444.43	460.00	5,100,000.00

Figure 4.6 functions as the visual representation of Table 4.8.

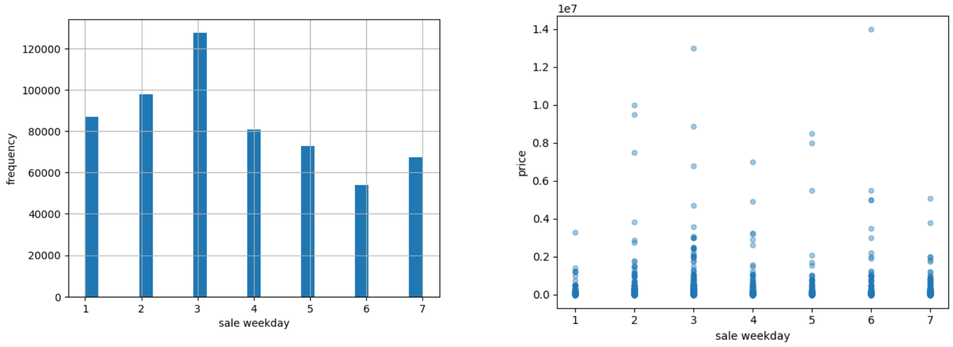


Figure 4.6: Histogram and Scatter Plot of Sale Weekday

4.3.2 Domain Name Related Features

This indicates domain name related features that can be extracted within (such as extension, domain length) or collected externally (such as extension popularity).

4.3.2.1 Name

A domain name is a combination of name and extension joined with a dot. So, this feature is interested in left side of this combination. Table 4.9 provides top 5 name; most popular one is poker with doc. count of 57, which is followed by casino, sex, casinos and de respectively. As it can be predicted, names containing words that belong to a certain category can be highly effective on the value.

Table 4.9: Top 5 Name with Stats

Name	Doc. Count (%)	Min	Mean	Median	Max
poker	57 (0.000)	210.00	50,761.46	1,670.00	1,000,000.00
casino	49 (0.000)	197.00	142,232.37	2,600.00	5,500,000.00
sex	41 (0.000)	100.00	751,057.12	2,509.00	14,000,000.00
casinos	30 (0.000)	170.00	8,446.43	4,880.00	52,000.00
de	30 (0.000)	260.00	3,946.77	2,099.00	21,500.00

4.3.2.2 IDN Status

A domain name can be represented in two forms; punycode and unicode. Starting with 'xn--' the ultimate indication of it being an IDN domain in its puny-coded representation. Unicode representation enables domains being represented in their local languages. As indicated in Table 4.10, not being an IDN domain dramatically increases domain value.

Table 4.10: IDN Status with Stats

IDN Status	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
false	586,813 (0.999)	90.00	2,641.40	500.00	14,000,000.00
true	564 (0.001)	100.00	2,162.59	1,160.00	100,749.00

4.3.2.3 Extension

A domain name is a combination of name and extension joined with a dot. So, this feature is interested in right side of this combination. There is a total of 466 extension exists in sale dataset. Table 4.11 indicates that com is the most popular extension and dominates the sales with 76% percentage. However, it should be noted that its median value is quite low even-though max priced sale is labeled with this extension. Another point that should attract our attention is the higher median value of sales with country extensions (such as de and uk).

Table 4.11: Top 5 Extension with Stats

Extension	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
com	445,316 (0.758)	100.00	2,762.01	450.00	14,000,000.00
net	37,497 (0.064)	90.00	1,745.32	700.00	500,000.00
org	29,445 (0.050)	90.00	1,330.31	380.00	1,000,000.00
de	12,990 (0.022)	100.00	3,546.66	1,494.50	1,169,175.00
uk	5,407 (0.009)	100.00	4,823.31	1,832.00	1,099,798.00

4.3.2.4 Extension Type

An extension has to be member of whether gTLD, newgTLD or ccTLD. gTLD stands for generic top-level domain and includes extensions such as com, net, org, info, biz, etc.

newgTLD stands for new generic top-level domain and includes extensions such xyz, loan, top, club, vip, etc. ccTLD stands for country code top-level domain and includes all two-letter top-level domains such as com.tr, co.uk, co.in, etc. Majority of extension type information for the sales in dataset is gTLD with 90% percentage. It is followed by ccTLD with 10% and surprisingly no newgTLD exists in the dataset as provided in Table 4.12. Interestingly, max priced sale is with gTLD but median value is quite higher for the ones with ccTLD.

Table 4.12: Extension Types with Stats

Ext. Type	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
ccTLD	61,159 (0.104)	100.00	2,915.43	1,197.00	1,169,175.00
gTLD	526,218 (0.896)	90.00	2,609.03	459.00	14,000,000.00

4.3.2.5 Extension Popularity

Popularity of each extension is different. Especially extensions in gTLD category are more popular than the others in terms of their registered domain count. Extension popularity information for each extension is collected from Extension Information API of DOFO. Table 4.13 is similar to Table 4.11 since it shows the popularity for top 5 extension.

Table 4.13: Top 5 Extension Popularity with Stats

Ext. Popularity	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
144,000,788 (com)	445,316 (0.758)	100.00	2,762.01	450.00	14,000,000.00
13,963,995 (net)	37,497 (0.064)	90.00	1,745.32	700.00	500,000.00
10,418,263 (org)	29,445 (0.050)	90.00	1,330.31	380.00	1,000,000.00
10,091,393 (de)	12,990 (0.022)	100.00	3,546.66	1,494.50	1,169,175.00
10,061,688 (uk)	5,407 (0.009)	100.00	4,823.31	1,832.00	1,099,798.00

As it is observable in Figure 4.7, com is the most popular extension in terms of both sales and registered domain count.

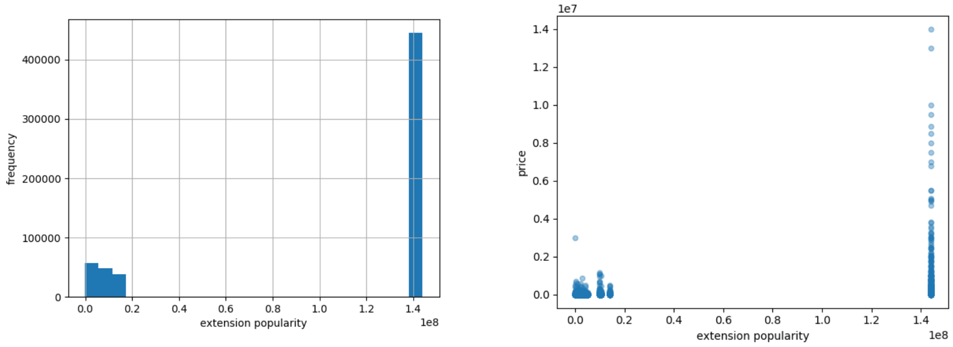


Figure 4.7: Histogram and Scatter Plot of Extension Popularity

4.3.2.6 Domain Length

Length of a domain name is calculated based on its name attribute being represented in punycode, which can be 63 at most. A total of 42 different lengths exists in sale dataset. Table 4.14 provides a sample of domain lengths sorted by median. As it is seen, median value is higher where domain length is lower.

Table 4.14: Domain Length with Stats Sample

Domain Len.	Doc. Count (%)	Min	Mean	Median	Max
1	683 (0.001)	100.00	17,554.24	1,827.00	6,784,000.00
2	3,649 (0.006)	100.00	24,881.15	1,450.00	8,500,000.00
3	27,969 (0.048)	100.00	6,137.49	711.00	14,000,000.00
8	44,035 (0.075)	100.00	2,760.32	704.00	3,200,000.00
9	45,287 (0.077)	100.00	2,415.95	681.00	3,250,000.00

Both Table 4.14 and Figure 4.8 indicate that value of domain name increases while domain length decreases.

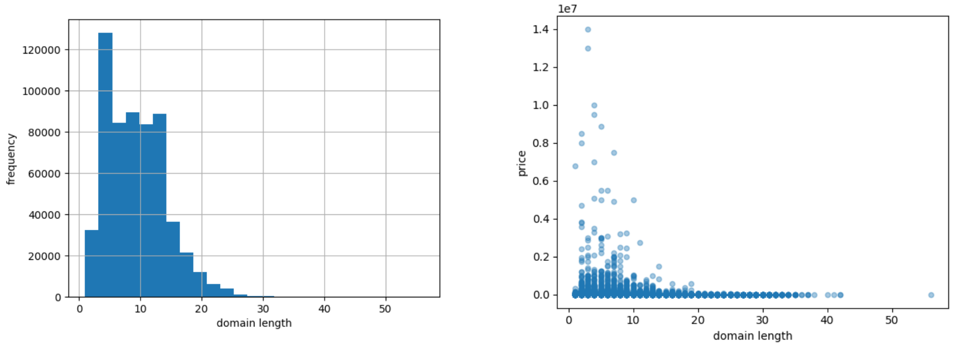


Figure 4.8: Histogram and Scatter Plot of Domain Length

4.3.2.7 Contains Letter

This feature indicates whether a domain contains at least one letter in its uni-coded representation. Domain names containing letters are majority with 96% percent and also more valuable according to statistics in Table 4.15 considering the median.

Table 4.15: Contains Letter with Stats

Cont. Letter	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
true	561,175 (0.955)	90.00	2,667.01	506.00	14,000,000.00
false	26,202 (0.045)	100.00	2,082.45	432.00	2,100,000.00

4.3.2.8 Letter Count

This feature indicates the number of letters a domain is containing in its uni-coded representation. Table 4.16 represents a sample of 5 letter count with their statistics sorted by median.

Table 4.16: A Sample of 5 Letter Count with Stats

Letter Count	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
7	38,001 (0.065)	100.00	3,742.88	760.00	7,500,000.00
3	22,153 (0.038)	100.00	6,653.78	756.00	14,000,000.00
6	34,768 (0.059)	90.00	3,585.55	750.00	5,500,000.00
8	43,695 (0.074)	100.00	2,763.89	709.00	3,200,000.00
9	44,972 (0.077)	100.00	2,420.13	678.50	3,250,000.00

It can be figured out from Figure 4.9, value is higher for domain names where letter count is lower.

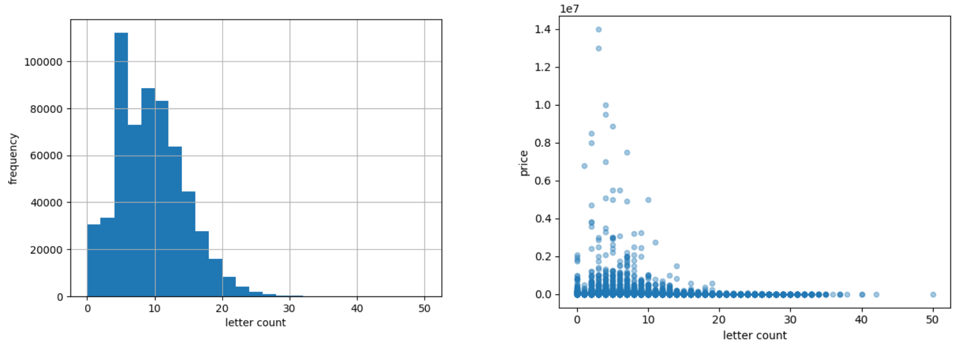


Figure 4.9: Histogram and Scatter Plot of Letter Count

4.3.2.9 Contains Number

This feature indicates whether a domain contains at least one digit in its uni-coded representation. Domain names not containing numbers are majority with 91% percent and also more valuable according to statistics in Table 4.17 considering the median.

Table 4.17: Contains Number with Stats

Cont. Number	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
false	536,153 (0.913)	90.00	2,727.18	510.00	14,000,000.00
true	51,224 (0.087)	100.00	1,738.24	410.00	2,100,000.00

4.3.2.10 Number Count

This feature indicates the number of digits a domain is containing in its uni-coded representation. Table 4.18 represents a sample of 5 digit count with their statistics sorted by median.

Table 4.18: A Sample of 5 Number Count with Stats

Number Count	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
5	11,168 (0.019)	100.00	950.77	550.00	245,000.00
0	536,153 (0.913)	90.00	2,727.18	510.00	14,000,000.00
1	11,836 (0.020)	100.00	1,857.61	500.00	800,000.00
2	6,573 (0.011)	100.00	2,886.52	470.00	1,960,800.00
3	5,406 (0.009)	100.00	4,181.51	353.00	2,100,000.00

It can be figured out from Figure 4.10, value is higher for domain names where number count is lower.

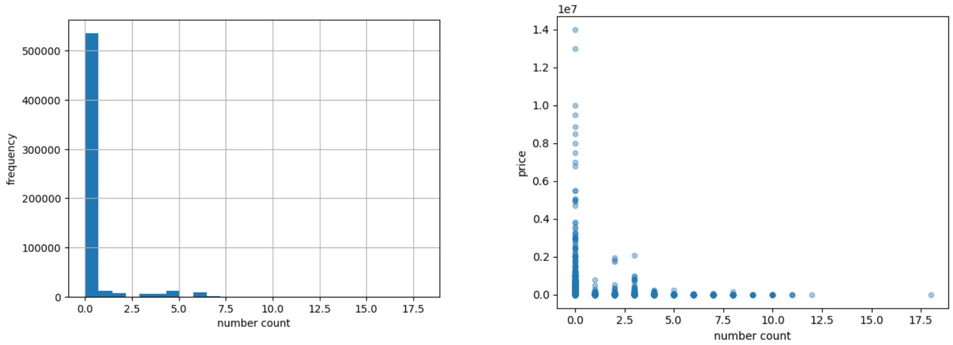


Figure 4.10: Histogram and Scatter Plot of Number Count

4.3.2.11 Contains Hyphen

This feature indicates whether a domain contains at least one hyphen in its uni-coded representation. Domain names not containing hyphens are majority with 97% percent and also slightly more valuable according to statistics in Table 4.19 considering the mean.

Table 4.19: Contains Hyphen with Stats

Cont. Hyphen	Doc. Count (%)	Min	Mean	Median	Max
false	571,516 (0.973)	90.00	2,678.91	500.00	14,000,000.00
true	15,861 (0.027)	100.00	1,272.77	499.00	209,916.00

4.3.2.12 Hyphen Count

This feature indicates the number of hyphens a domain is containing in its uni-coded representation. Table 4.20 represents a sample of 5 hyphen count with their statistics sorted by median.

Table 4.20: A Sample of 5 Hyphen Count with Stats

Hyphen Count	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
1	14,186 (0.024)	100.00	1,317.25	510.00	209,916.00
0	571,516 (0.973)	90.00	2,678.91	500.00	14,000,000.00
2	1,503 (0.003)	100.00	911.07	290.00	30,350.00
3	154 (0.000)	100.00	808.55	290.00	15,601.00
4	11 (0.000)	110.00	520.09	154.00	3,500.00

It can be figured out from Figure 4.11, value is higher for domain names where hyphen count is lower.

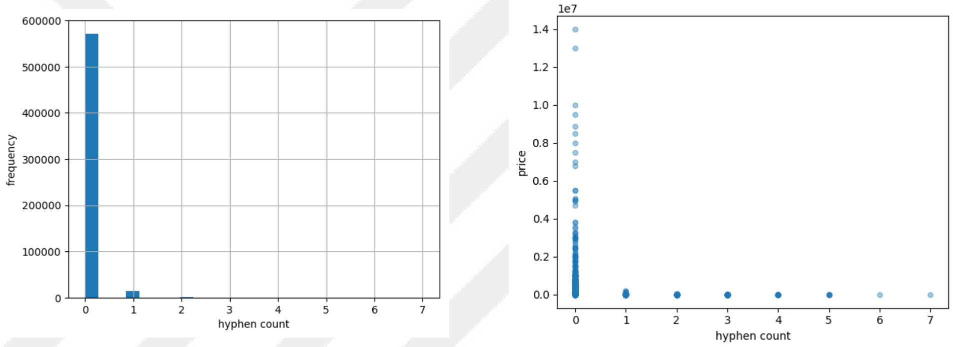


Figure 4.11: Histogram and Scatter Plot of Hyphen Count

4.3.2.13 Contains Emoji

This feature indicates whether a domain contains at least one emoji in its uni-coded representation. Domain names not containing emojis are majority with almost 100% percent and also more valuable according to statistics in Table 4.21 considering the median.

Table 4.21: Contains Emoji with Stats

Cont. Emoji	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
false	587,322 (1.000)	90.00	2,641.09	500.00	14,000,000.00
true	55 (0.000)	100.00	944.89	194.00	13,600.00

4.3.2.14 Emoji Count

This feature indicates the number of emoji a domain is containing in its uni-coded representation. Table 4.22 represents emoji counts with their statistics sorted by median.

Table 4.22: Emoji Counts with Stats

Emoji Count	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
0	587,322 (1.000)	90.00	2,641.09	500.00	14,000,000.00
1	55 (0.000)	100.00	944.89	194.00	13,600.00

It can be figured out from Figure 4.12, value is higher for domain names where emoji count is lower.

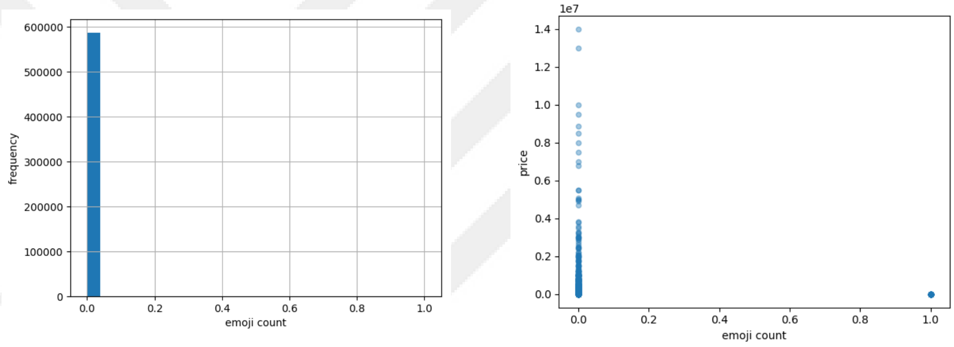


Figure 4.12: Histogram and Scatter Plot of Emoji Count

4.3.3 Extension Stats Related Features

Name property of a domain name is highly determinative in terms of valuation. So, knowing number of extensions available, registered and forsale for a given name will provide additional information in order to understand characteristic of that given name. For that very reason, we used DOFO's Extension Stats API and collected available, registered and forsale extension count data for each name in the sale dataset.

4.3.3.1 Available Extension Count

Figure 4.13 indicates that available extension count for most of the name are high and value for a domain name increases while available extension count increases.

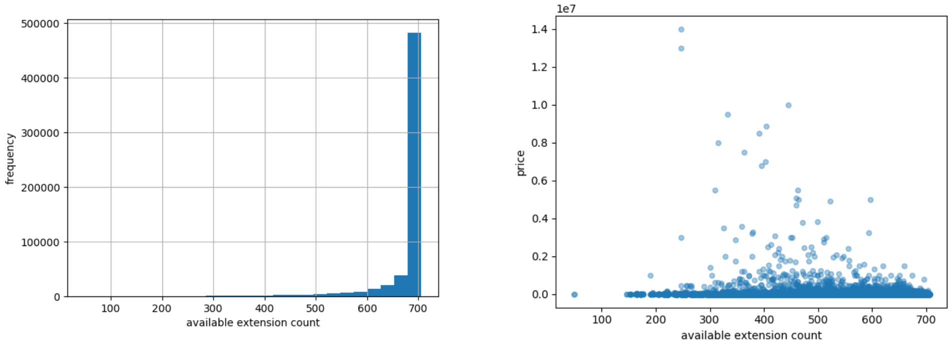


Figure 4.13: Histogram and Scatter Plot of Available Extension Count

4.3.3.2 Forsale Extension Count

Figure 4.14 indicates that forsale extension count for most of the name are low and value for a domain name increases while forsale extension count decreases.

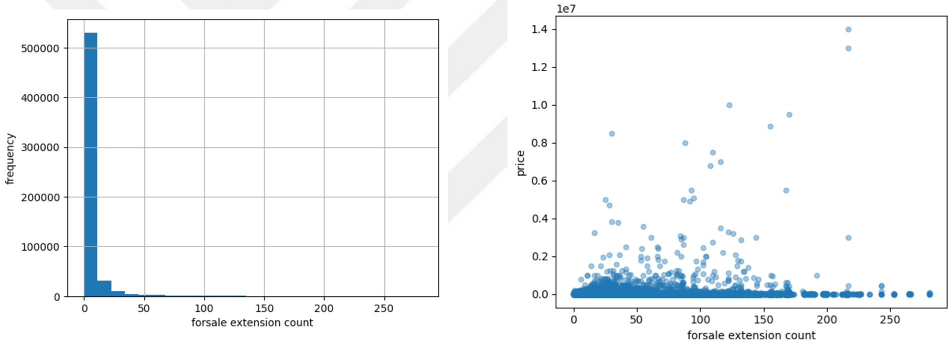


Figure 4.14: Histogram and Scatter Plot of Forsale Extension Count

4.3.3.3 Registered Extension Count

Figure 4.15 indicates that registered extension count for most of the name are low and value for a domain name increases while registered extension count decreases.

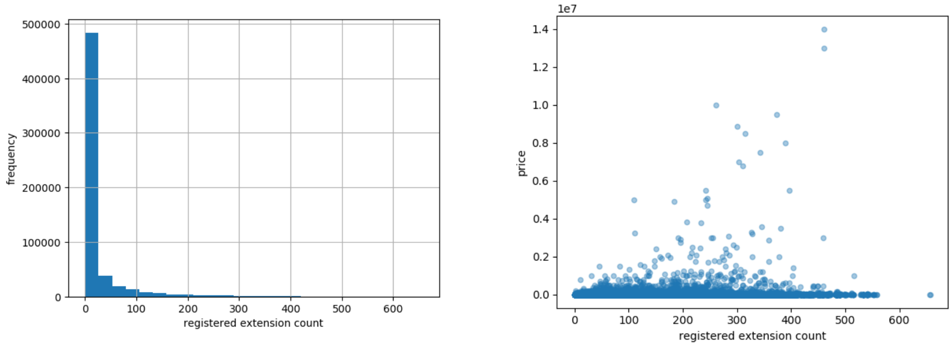


FIGURE 4.15: Histogram and Scatter Plot of Registered Extension Count

4.3.4 DNLD Related Features

DNLD is very important in terms of understanding characteristics of a domain name since it enables us to identify its language, extracts keywords within and derive various new features on top of them as mentioned in Chapter 3. So, below is the features extracted based on language identification (LI) and word segmentation sub-studies. Details on technical side of DNLD and how successful it is will be presented in Chapter 5 and Chapter 6 respectively.

4.3.4.1 Domain Language

Domain names in sale dataset are labeled with 40 different languages as a result of our DNLD study, even-thought it normally support much more than that. Table 4.23 indicates that language of majority of domain names are English, is it may not be so much decisive in terms of valuation within the group, but it is quite disjunctive compared to other languages since domain names with language English seems to be slightly more valuable. On the other hand, mean value of German is not as popular as English but its median value is higher. Normally domain names with no language assigned are labelled with "-". It is interesting to note that, it is ranked as 3 in terms of popularity. Domain names in this category are mostly combination of numbers such as 855.com, 35.org and 02043.com or combination of unmeaningful chars and digits such as f-h-s.com, z-n.com, m4x.com, g4b.com, 121s.com and x360.com.

Table 4.23: Top 5 Domain Language with Stats

Domain Lang.	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
English (en)	403,844 (0.688)	90.00	2,822.84	525.00	14,000,000.00
German (de)	39,219 (0.067)	100.00	2,831.91	803.00	8,888,888.00
-	32,899 (0.056)	100.00	2,306.80	436.00	6,784,000.00
French (fr)	12,594 (0.021)	100.00	2,424.72	560.00	1,058,830.00
Spanish (es)	10,412 (0.018)	100.00	2,830.52	582.50	2,100,000.00

4.3.4.2 Keywords

A total of 127,712 different keywords extracted from domain names in the sale dataset. Table 4.24 contains statistics of top 20 keywords extracted during DNLD process. Popularity of a keyword (in terms of count) does not necessarily indicate for that domain being more valuable; it has impact but additional features are needed to uncover the real accurate impact. So, following features are generated while keeping this in mind.

Table 4.24: Top 20 Keyword with Stats

Keyword	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
the	8,102 (0.014)	100.00	1,270.20	365.00	500,000.00
online	4,043 (0.007)	100.00	1,972.72	600.00	165,000.00
my	3,117 (0.005)	100.00	2,569.31	726.00	1,200,000.00
home	2,395 (0.004)	100.00	1,850.29	585.00	166,650.00
of	2,376 (0.004)	100.00	1,116.20	340.50	45,000.00
web	2,330 (0.004)	100.00	1,668.64	576.00	150,000.00
and	2,181 (0.004)	100.00	1,043.73	310.00	28,888.00
free	2,102 (0.004)	100.00	3,026.13	530.00	500,000.00
for	1,976 (0.003)	100.00	1,695.20	349.50	500,000.00
shop	1,959 (0.003)	100.00	3,900.13	1,000.00	3,500,000.00
world	1,953 (0.003)	100.00	2,587.72	666.00	1,200,000.00
health	1,932 (0.003)	100.00	1,717.74	617.00	55,537.00
group	1,861 (0.003)	100.00	1,436.58	698.00	50,000.00
life	1,803 (0.003)	100.00	1,883.93	750.00	171,750.00
media	1,722 (0.003)	100.00	1,413.62	535.00	60,000.00
news	1,615 (0.003)	100.00	1,832.93	355.00	200,000.00
auto	1,597 (0.003)	90.00	2,711.36	650.00	440,000.00
insurance	1,567 (0.003)	100.00	3,136.09	895.00	570,000.00
net	1,541 (0.003)	100.00	1,953.31	700.00	100,000.00
design	1,529 (0.003)	100.00	1,233.09	407.00	40,000.00

4.3.4.3 Number of Keywords

This feature indicates total number of keywords that a domain name contains. This value has a scale from 0 to 9 and as it is stated in Table 4.25, most popular keyword count is 2 with 57%, which is followed by 1 (27%), 3 (10%), 0 (6%) and 4 (1%). It is noticeable that domain names with keyword count 1 are the most valuable ones. Example for 2 can be kidsong.com, where as waystomakemoneyontheinternet.com can be counted for 7.

TABLE 4.25: Number of Keywords with Stats

N. of Key.	Doc. Count (%)	Min	Mean	Median	Max
0	32,899 (0.056)	100.00	2,306.80	436.00	6,784,000.00
1	155,787 (0.265)	90.00	5,654.60	812.00	14,000,000.00
2	331,995 (0.565)	90.00	1,587.00	491.00	1,500,000.00
3	57,969 (0.099)	100.00	1,046.69	293.00	500,000.00
4	7,657 (0.013)	100.00	825.56	220.00	110,000.00
5	919 (0.002)	100.00	496.22	205.00	10,000.00
6	121 (0.000)	100.00	572.14	201.00	5,000.00
7	25 (0.000)	100.00	570.92	186.00	2,850.00
8	4 (0.000)	110.00	889.50	912.50	1,623.00
9	1 (0.000)	2,555.00	2,555.00	2,555.00	2,555.00

Figure 4.16 functions as the visual representation of Table 4.25 and indicates that value is higher for domain names that having less keywords.

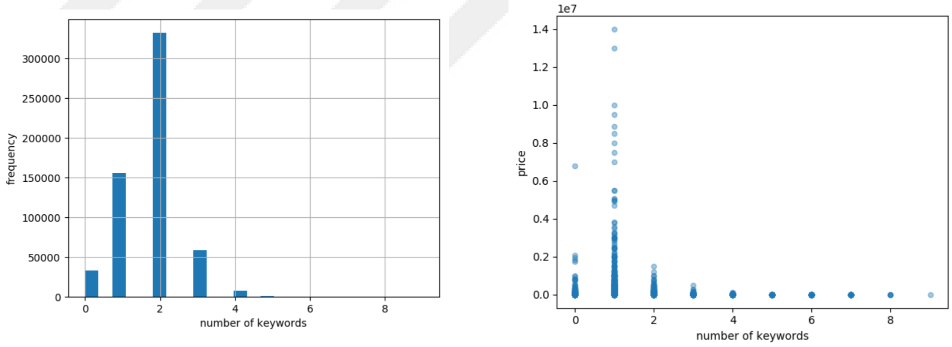


FIGURE 4.16: Histogram and Scatter Plot of Number of Keywords

4.3.4.4 Meaningful Char Count

This feature indicates total number of chars being a part of a keyword within domain name. Table 4.26 represents a sample of 5 meaningful char count with stats sorted by median.

Table 4.26: A Sample of 5 Meaningful Char Count with Stats

M. Char Count	Doc. Count (%)	Min	Mean	Median	Max
6	34,493 (0.059)	90.00	3,611.10	750.00	5,500,000.00
7	37,706 (0.064)	100.00	3,746.40	750.00	7,500,000.00
8	43,420 (0.074)	100.00	2,741.54	709.00	3,200,000.00
9	44,806 (0.076)	100.00	2,422.86	676.00	3,250,000.00
5	30,355 (0.052)	100.00	4,643.29	671.00	8,888,888.00

Figure 4.17 functions as the visual representation of Table 4.26. As it is seen, value of a domain name increases while meaningful char count decreases.

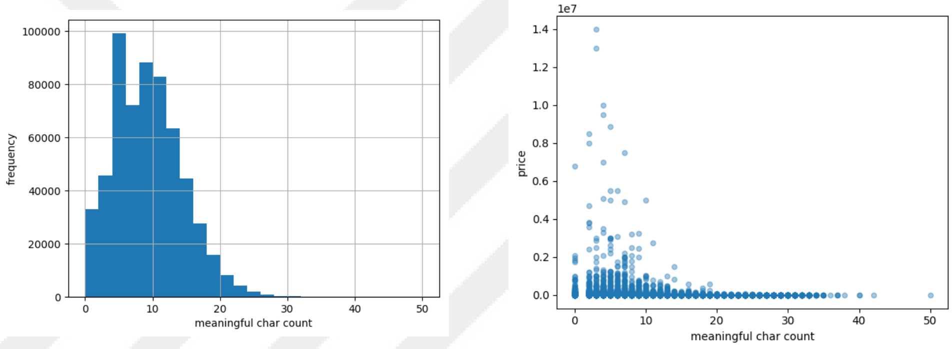


Figure 4.17: Histogram and Scatter Plot of Meaningful Char Count

4.3.4.5 Unmeaningful Char Count

This feature indicates total number of chars not being a part of a keyword within domain name. Table 4.27 represents a sample of 5 unmeaningful char count with stats sorted by median.

Table 4.27: A Sample of 5 Unmeaningful Char Count with Stats

U. Char Count	Doc. Count (%)	Min	Mean	Median	Max
5	11,524 (0.020)	100.00	950.14	550.00	245,000.00
0	498,259 (0.848)	90.00	2,818.78	520.00	14,000,000.00
2	8,625 (0.015)	100.00	2,943.62	418.00	1,960,800.00
1	42,422 (0.072)	100.00	1,579.63	409.00	6,784,000.00
3	9,611 (0.016)	100.00	2,904.52	409.00	2,100,000.00

Figure 4.18 functions as the visual representation of Table 4.27. As it is seen, value of a domain name increases while unmeaningful char count decreases.

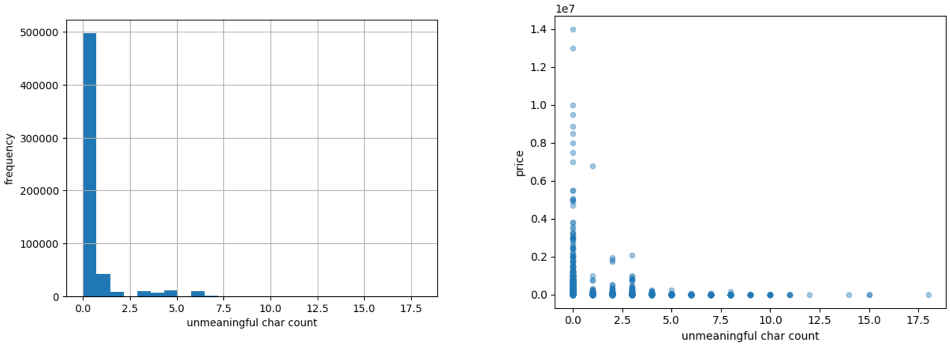


FIGURE 4.18: Histogram and Scatter Plot of Unmeaningful Char Count

4.3.4.6 Meaningfulness Ratio

This feature indicates total number of chars being a part of a keyword within domain name divided by domain length. Table 4.28 represents a sample of 5 meaningfulness ratio with stats which is first rounded by 0.05 and then sorted by median.

TABLE 4.28: A Sample of 5 Meaningfulness Ratio with Stats

M. Ratio	Doc. Count (%)	Min	Mean	Median	Max
0.65	4,393 (0.007)	100.00	1,461.08	645.00	60,001.00
0.85	5,054 (0.009)	100.00	2,025.66	560.00	750,000.00
0.80	3,427 (0.006)	100.00	1,843.45	560.00	75,000.00
1.00	498,260 (0.848)	90.00	2,818.78	520.00	14,000,000.00
0.90	12,810 (0.022)	100.00	1,494.01	464.00	1,000,000.00

Figure 4.19 functions as the visual representation of Table 4.28. As it is seen, domain names with high meaningfulness ratio are more valuable than the others.

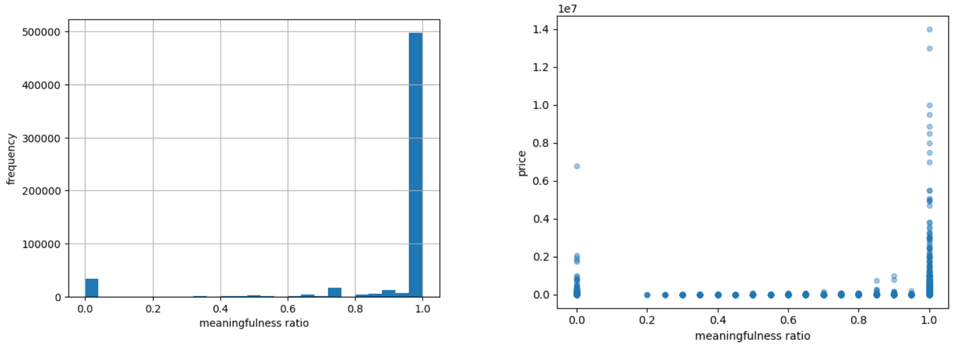


FIGURE 4.19: Histogram and Scatter Plot of Meaningfulness Percentage

4.3.4.7 Has Unassigned Char

This feature indicates whether a domain contains at least one unassigned char in terms of not being a part of a keyword in its uni-coded representation. Domain names not having unassigned char are majority with 85% percent and also more valuable according to statistics in Table 4.29 considering median values.

TABLE 4.29: Has Unassigned Char with Stats

Has Un. Char	**Doc. Count (%)**	**Min**	**Mean**	**Median**	**Max**
false	498,259 (0.848)	90.00	2,818.78	520.00	14,000,000.00
true	89,118 (0.152)	100.00	1,646.61	400.00	6,784,000.00

4.3.4.8 Keyword Length

We have experimented that value of a domain name increases as the length of it decreases. We wonder if this is the case for the keywords contained by a domain name as well. So, three different versions of keyword length are generated and named as min-mean-max as a domain name may contain more than one keyword.

4.3.4.8.1 Min. Keyword Length

This feature indicates the length of the shortest word among the words that a domain name contains. If no word is extracted, set to 0. Table 4.30 represents top 5 min. keyword length with stats.

Table 4.30: Top 5 Min. Keyword Length with Stats

Min. Key. Len.	Doc. Count (%)	Min	Mean	Median	Max
3	127,611 (0.217)	100.00	2,293.79	455.00	14,000,000.00
4	126,900 (0.216)	90.00	2,482.36	550.00	9,999,950.00
2	99,340 (0.169)	100.00	1,770.57	309.00	8,500,000.00
5	77,569 (0.132)	100.00	2,950.07	616.00	8,888,888.00
6	51,353 (0.087)	90.00	3,022.66	750.00	5,500,000.00

Figure 4.20 functions as the visual representation of Table 4.30. As it is observable, domain value increases where min. keyword length decreases.

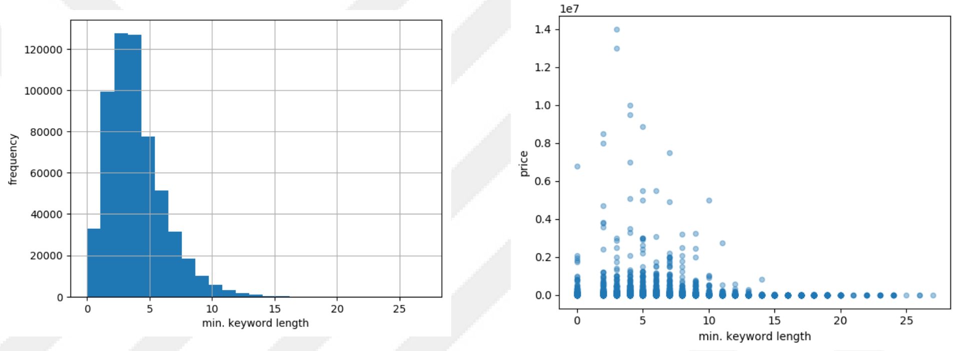


Figure 4.20: Histogram and Scatter Plot of Min. Keyword Length

4.3.4.8.2 Avg. Keyword Length

This feature indicates total length of keywords within a domain name divided by keyword count. If no word is extracted, set to 0. Table 4.31 represents top 5 avg. keyword length with stats.

Table 4.31: Top 5 Avg. Keyword Length with Stats

Avg. Key. Len.	Doc. Count (%)	Min	Mean	Median	Max
4.0	71,444 (0.122)	100.00	3,072.49	580.00	9,999,950.00
5.0	60,559 (0.103)	100.00	3,290.57	660.00	8,888,888.00
3.0	56,325 (0.096)	100.00	3,330.43	500.00	14,000,000.00
6.0	48,237 (0.082)	90.00	3,120.53	667.00	5,500,000.00
2.0	44,843 (0.076)	100.00	2,259.89	280.00	8,500,000.00

Figure 4.21 functions as the visual representation of Table 4.31. As it is seen, domain names with low avg. keyword length are more valuable than the others.

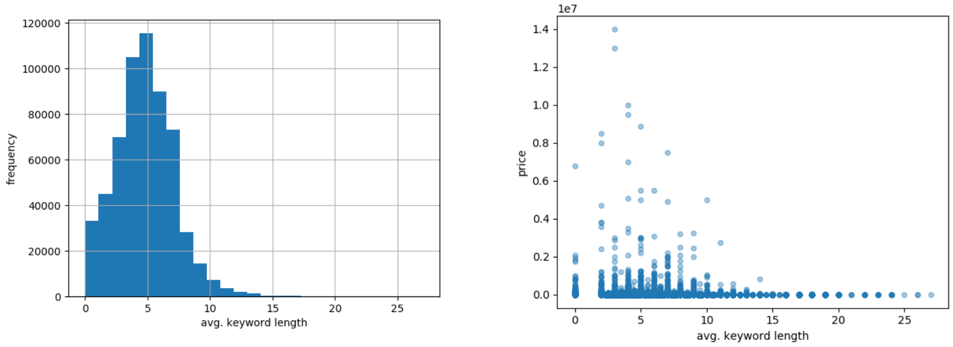


FIGURE 4.21: Histogram and Scatter Plot of Avg. Keyword Length

4.3.4.8.3 Max. Keyword Length

This feature indicates the length of the longest word among the words of the domain name. If no word is extracted, set to 0. Table 4.32 represents top 5 max. keyword length with stats.

TABLE 4.32: Top 5 Max. Keyword Length with Stats

Max. Key. Len.	Doc. Count (%)	Min	Mean	Median	Max
6	83,452 (0.142)	90.00	2,479.35	579.00	5,500,000.00
5	81,728 (0.139)	100.00	2,865.60	610.00	8,888,888.00
4	79,130 (0.135)	100.00	2,947.78	577.00	9,999,950.00
7	72,854 (0.124)	90.00	2,619.16	532.00	7,500,000.00
3	60,239 (0.103)	100.00	3,135.97	446.00	14,000,000.00

Figure 4.22 functions as the visual representation of Table 4.32. As it is observable, domain value increases where max. keyword length decreases.

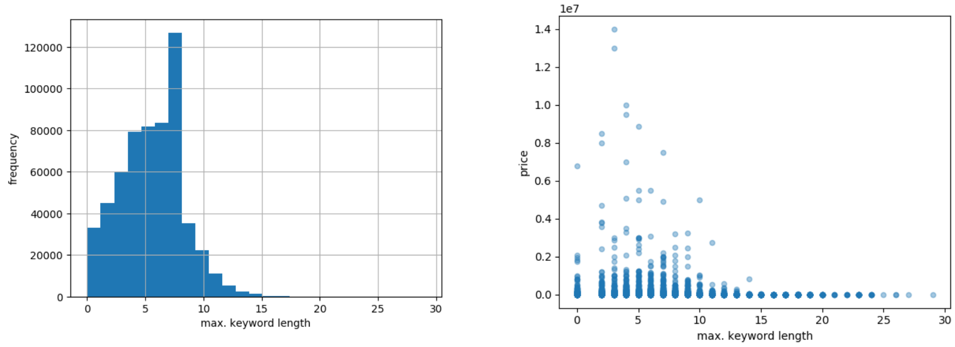


FIGURE 4.22: Histogram and Scatter Plot of Max. Keyword Length

4.3.4.9 Meaningful Language Count

This feature indicates number of languages that each keyword within a domain name are meaningful in. Table 4.33 represents a sample of 5 meaningful language count with stats sorted by median.

TABLE 4.33: A Sample of 5 Meaningful Language Count with Stats

M. Lang. Count	Doc. Count (%)	Min	Mean	Median	Max
9	4,664 (0.008)	100.00	2,177.19	635.00	502,225.00
4	7,382 (0.013)	100.00	2,181.60	625.00	400,000.00
6	6,202 (0.011)	100.00	2,227.73	625.00	600,000.00
5	6,678 (0.011)	100.00	2,278.51	613.50	1,350,000.00
7	5,692 (0.010)	100.00	2,151.19	600.00	500,000.00

Figure 4.23 functions as the visual representation of Table 4.33. As it is seen, domain names with higher meaningful language count are more valuable than the others.

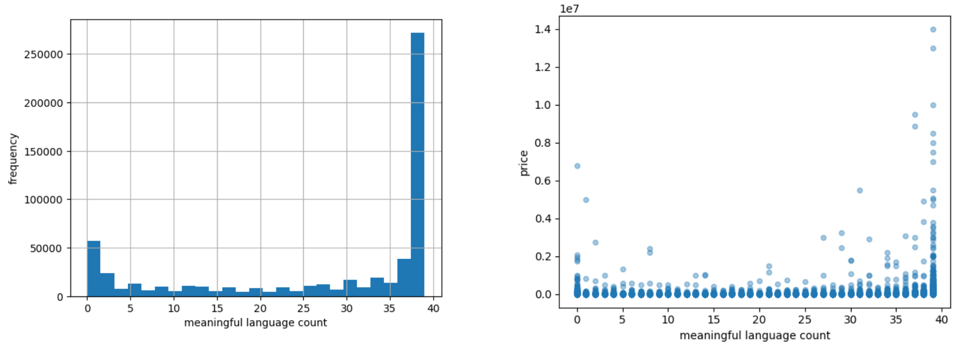


FIGURE 4.23: Histogram and Scatter Plot of Meaningful Language Count

4.3.4.10 Partially Meaningful Language Count

This feature indicates number of languages that at least one keyword within a domain name are meaningful in. Table 4.34 represents a sample of 5 partially meaningful language count with stats sorted by median.

Table 4.34: A Sample of 5 Partially Meaningful Language Count with Stats

P. M. Lang.	Doc. Count (%)	Min	Mean	Median	Max
9	2,489 (0.004)	100.00	3,148.01	1,032.00	502,225.00
6	3,314 (0.006)	100.00	2,911.41	1,003.00	240,000.00
21	1,801 (0.003)	100.00	5,293.21	1,000.00	1,500,000.00
22	1,735 (0.003)	100.00	3,276.86	1,000.00	550,000.00
8	2,693 (0.005)	100.00	4,740.92	1,000.00	2,430,000.00

Figure 4.24 functions as the visual representation of Table 4.34. As it is seen, domain names with higher partially meaningful language count are more valuable than the others.

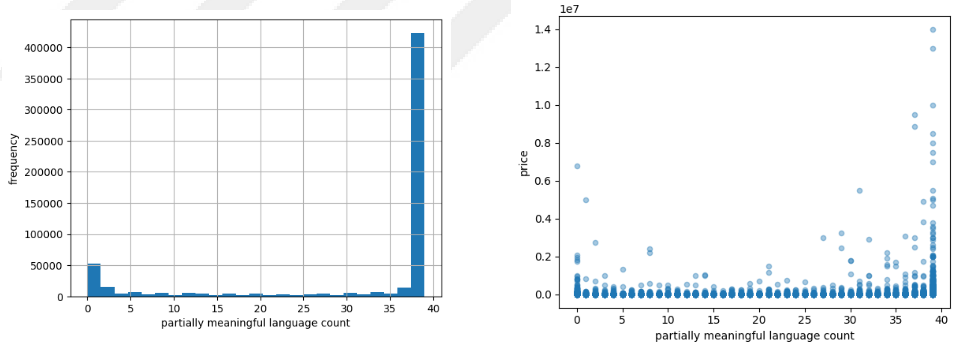


Figure 4.24: Histogram and Scatter Plot of Partially Meaningful Language Count

4.3.4.11 Keyword Popularity

We have seen that value of a domain name increases as the extension popularity increases. We wonder if this is the case for the keywords in a domain name as well. The popularity of a keyword is computed as word count divided by word rank for the language that the keyword belongs to. So, three different versions of keyword popularity are generated and named as min-mean-max as a domain name may contain more than one keyword.

4.3.4.11.1 Min. Keyword Popularity

This feature indicates the popularity of the least popular of words that a domain name contains. As it can be observed in Figure 4.25, domain names with higher min. keyword popularity are more valuable than the others.

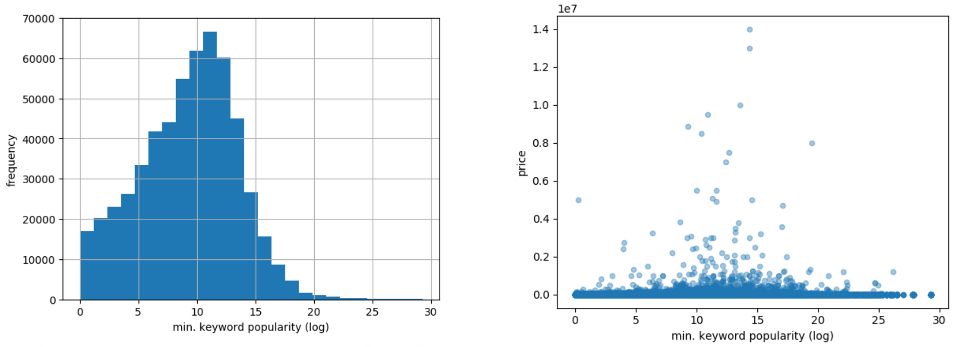


FIGURE 4.25: Histogram and Scatter Plot of Min. Keyword Popularity in Log Scale

4.3.4.11.2 Avg. Keyword Popularity

This feature is computed as average of the popularity of words that a domain name contains. As it can be observed in Figure 4.26, domain names with higher avg. keyword popularity are more valuable than the others.

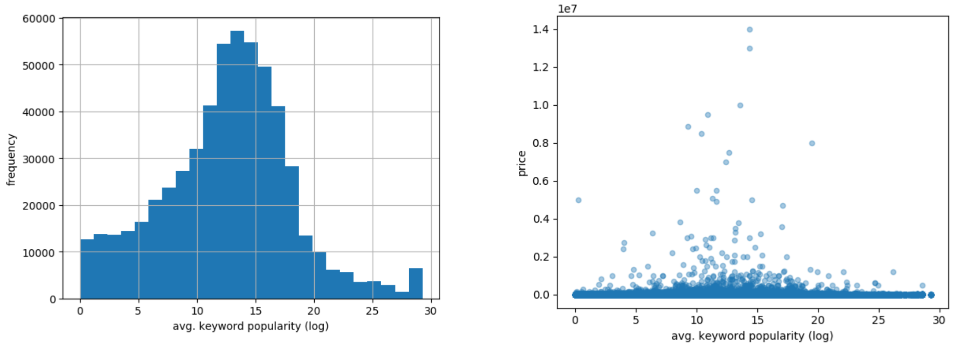


FIGURE 4.26: Histogram and Scatter Plot of Avg. Keyword Popularity in Log Scale

4.3.4.11.3 Max. Keyword Popularity

This feature indicates the popularity of the most popular of words that a domain name contains. As it can be observed in Figure 4.27, domain names with higher max. keyword popularity are more valuable than the others.

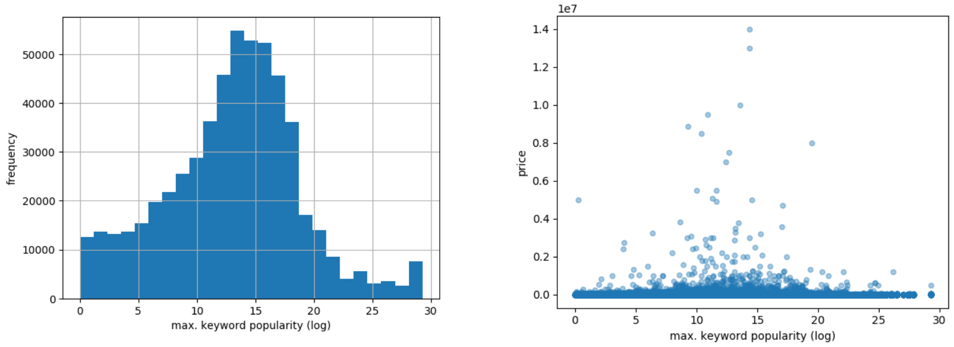


Figure 4.27: Histogram and Scatter Plot of Max. Keyword Popularity in Log Scale

4.4 Column Transformation

4.4.1 Numeric Feature Transformation

Numeric feature transformation consists of standard scaling only and applied on following features:

- Extension popularity
- Domain length
- Letter count
- Number count
- Hyphen count
- Emoji count
- Available extension count
- Forsale extension count
- Registered extension count
- Number of keywords
- Meaningful char count
- Unmeaningful char count
- Meaningfulness ratio
- Keyword length (min-avg-max)
- Meaningful language count
- Partially meaningful language count
- Keyword popularity (min-avg-max)

4.4.2 Categorical Feature Transformation

Categorical feature transformation consists of one-hot encoding only and applied on following features:

- Venue
- Sale month
- Sale season
- Sale weekday
- IDN status
- Extension
- Extension type
- Contains letter
- Contains number
- Contains hyphen
- Contains emoji
- Domain language
- Has unassigned char

4.4.3 Text Feature Transformation

Text feature transformation consists of count vectoring only and applied on following features:

- Keywords

4.4.4 Ngram Representation

Ngram representation consists 3-gram generation followed by TF-IDF vectoring and applied on following features:

- Name

Chapter 5

Methodology & Modeling

5.1 Domain Name Language Detection (DNLD)

Language identification (LI) is classified as a Natural Language Processing (NLP) task and it is one of the well studied topics within the field. Although there is a number of studies done on it already, these studies start with the assumption of having relatively long text as input. Since the scope of this study is based on domain names, which are limited with *63 unseparated characters* at most, the approaches that they adopted cannot be directly applicable in this case. Challenge in DNLD is not only detecting the language but also concurrently identifying the words that the domain name is composed of (referred to as keyword extraction). DNLD is very important in terms of understanding characteristics of a domain name since it enables us to identify its language, extracts keywords within and derive various new features on top of them, so it is essential to the completeness of the study. We model this challenging problem as a weighted interval scheduling problem and develop a dynamic programming based solution.

In weighted interval scheduling, each interval is assigned a certain value (i.e. weight), and the objective is basically to come up with the set of non-overlapping intervals that provides the maximum total weight. Each interval is an individual word, referred to as keyword, in our domain name language identification and keyword extraction problem. A greedy algorithm that simply maximizes the number of intervals (i.e. keywords) may not necessarily lead to a better solution as the words with shorter lengths are common across multiple languages. Similarly, the usage frequency (i.e. popularity) of a word in a language is also a good proxy in language and keyword selection. Subsequently, a number of keyword weighting scheme that takes into account word length and popularity is proposed.

A weighted interval scheduling algorithm implementation [60] that uses binary search to identify the non-overlapping intervals are utilized in this work. Since the number of possible intervals in domain names can be quite large in some cases, a dynamic programming based implementation that uses memorization is adopted. Time complexity of this implementation is $\mathcal{O}(n \log n)$.

We propose a number of weighting schemes to decide on language and extract keywords within a domain name by addressing the need for ranking among various possible keyword segmentation of a domain name in multiple-languages. Weighting approach takes into account word length and word's usage frequency (i.e. word popularity). Since we do not have access to the usage frequency of each word in any given language, word rank (i.e. usage rank of the word) and word count (i.e. total number of words) in a language are utilized instead to quantify the word popularity. A number of proposed weighting schemes are provided below:

1. $word\_len * log((\frac{word\_rank^2}{word\_count*(word\_count+1)*(2*word\_count+1)/6})^{-1})$
2. $word\_len * log(word\_len) * log((\frac{word\_rank^2}{word\_count*(word\_count+1)*(2*word\_count+1)/6})^{-1})$
3. $word\_len^2 * log((\frac{word\_rank^2}{word\_count*(word\_count+1)*(2*word\_count+1)/6})^{-1})$
4. $word\_len * log((\frac{word\_rank}{word\_count})^{-1})$
5. $word\_len * log(word\_len) * log((\frac{word\_rank}{word\_count})^{-1})$
6. $word\_len^2 * log((\frac{word\_rank}{word\_count})^{-1})$
7. $word\_len * log((\frac{word\_rank}{word\_count})^{-2})$
8. $word\_len * log(word\_len) * log((\frac{word\_rank}{word\_count})^{-2})$
9. $word\_len^2 * log((\frac{word\_rank}{word\_count})^{-2})$
10. $log((\frac{word\_rank^2}{word\_count*(word\_count+1)*(2*word\_count+1)/6})^{-1})$
11. $log((\frac{word\_rank}{word\_count})^{-1})$

12. $log((\frac{word\_rank}{word\_count})^{-2})$

Proposed weighting approaches fundamentally consists of two parts. First part calibrates the effect of word length in the weight of word and similarly second part calibrates the effect of word's popularity in the weight calculation. The weighting schemes listed above vary the effect of these two parts in weight calculation. Weighting schemes 10-12 only consider word popularity without paying any attention to the word lengths. Experimental evaluation of these weighting schemes are shared in Chapter 6.

Language identification and keyword extraction algorithm (DNLD) fundamentally generates ngrams from domain name and look them up in Fasttext dataset for each available language. If an ngram exists in a language, correspond weight is calculated. Finally weighted interval scheduling algorithm is run for each language with the identified keywords and the language and keyword combination with the highest total weight is selected. Algorithm 1 shows the pseudo-code for DNLD Algorithm. Firstly, ngrams based on name part of given domain name are generated (lines 1 to 2). Secondly, generated ngrams are checked against word look-up dataset prepared at data wrangling phase as explained in 4.2.1. And the ones exist at least in one of the languages are stored with their meta (starting & ending index) and weight information by language (lines 3 to 16). Thirdly, language based intervals are fed into weighted interval scheduling algorithm separately (lines 17 to 21). And lastly, domain name's language and keywords within are decided based on maximum total weight of best schedule candidate retrieved by each language (lines 22 to 23).

Table 5.1: Top 5 word stats for *hello* with Fasttext-v1 and weight-model-3

Language	**Word Rank**	**Word Count in Language**	**Word Weight**
en	4,796	2,256,620	207.927
de	20,955	2,217,149	183.768
fr	8,181	1,073,866	181.404
ko	7,789	876,598	177.295
ru	22,336	1,727,980	176.723

Algorithm 1 DNLD

```
Require: A domain name dn
Ensure: The predicted (lang, keywords) pair for dn
name = NAME(dn);
ngrams = NGRAMS(name);
lang2intervals = Map<String, Array<Interval>>;
for each ngram ∈ ngrams do
    if ngram ∈ WORD_LOOKUP then
        word_length = LENGTH(ngram);
        langs = WORD_LOOKUP(ngram);
        for each lang ∈ langs do
            word_rank = WORD_LOOKUP(ngram, lang);
            word_count = LANG_LOOKUP(lang);
            weight = WEIGHT(word_length, word_rank, word_count);
            interval = INTERVAL(ngram, ngram_start, ngram_end, weight);
            lang2intervals[lang].add(interval);
        end for
    end if
end for
lang2schedule = Map<String, Schedule>>;
for each (lang, intervals) ∈ lang2intervals do
    schedule = SCHEDULE_WEIGHTED_INTERVALS(intervals);
    lang2schedule[lang] = schedule;
end for
best_schedule = MAX(lang2schedule);
return best_schedule;
```

Table 5.2: Top 5 word stats for *world* with Fasttext-v1 and weight-model-3

Language	Word Rank	Word Count in Language	Word Weight
en	88	2,256,620	272.275
nl	101	847,892	246.426
de	817	2,217,149	235.986
es	518	948,309	222.817
ru	1,438	1,727,980	220.869

Table 5.1 and Table 5.2 represent top 5 languages and related stats for *hello* and *world* respectively as examples, where word weight calculated based on weight model 3 and trained with Fasttext-v1. And Table 5.3 is shared to illustrate DNLD result for *helloworld.com*.

Table 5.3: Top 5 DNLD stats for *helloworld.com* with Fasttext-v1 and weight-model-3

Language	Keywords	Language Score
en	[(hello, 207.927), (world, 272.275)]	480.202
de	[(hello, 183.768), (world, 235.986)]	419.755
nl	[(hello, 168.050), (world, 246.426)]	414.476
ru	[(hello, 176.723), (world, 220.869)]	397.593
fr	[(hello, 181.404), (world, 214.249)]	395.653

5.1.1 DNLD as a Micro-service

Following architecture is established and DNLD is launched as a micro-service. Loosely speaking, we choose technologies in order to enable a system that has the following capabilities: being extremely fast, scalable, highly available, replicated, distributed, fault-tolerant and secure. More specifically, we choose Nginx to function as a load balancer so that demand can be distributed among workers, Hazelcast to store language and word look-up mappings explained in 4.2.1 and MongoDB to store DNLD results by name so that computation cost will be minimized by providing the capabilities mentioned above.

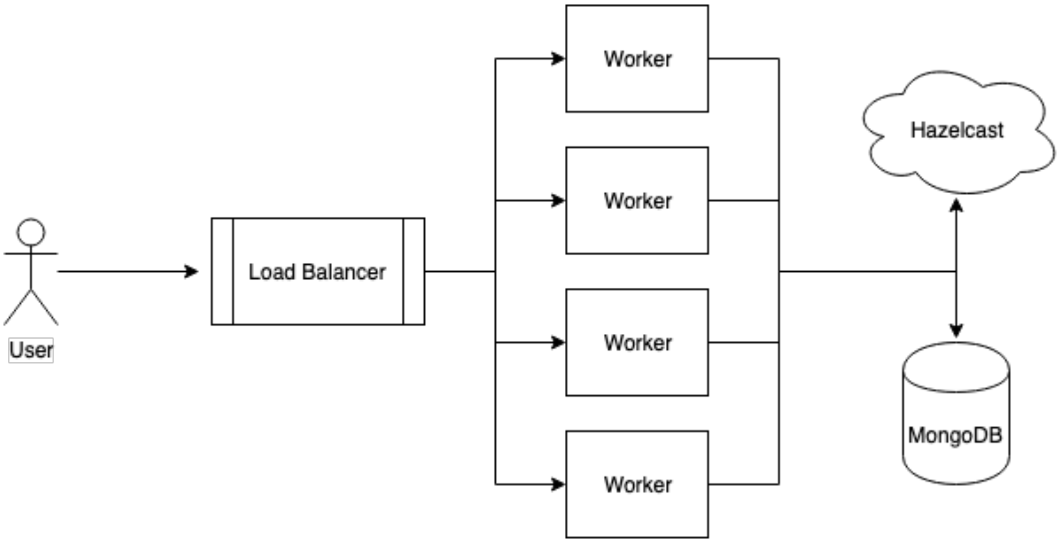


Figure 5.1: General Architecture of DNLD Micro-service

Algorithm 2 shows the pseudo-code for how DNLD API functions. Basically, when a request hits to a worker, it extracts the name (line 1) and checks cache (i.e. MongoDB) if it requested before (line 2), if so, then retrieve it (line 3), otherwise, compute DNLD with look-ups stored in Hazelcast and cache it (line 5 to 6), and finally return either retrieved or computed result (line 8).

Algorithm 2 DNLD API

```
Require: A domain name dn
Ensure: The predicted (lang, keywords) pair for dn
1: name = NAME(dn);
2: if name $\in$ CACHE then
3:     schedule = CACHE(name);
4: else
5:     schedule = DNLD(dn);
6:     CACHE[name] = schedule;
7: end if
8: return schedule;
```

5.2 Domain Name Price Estimation (DNPE)

The primary data that is used in domain name price estimation is historical data of domain sales, which is exported from DOFO's storage. An enrichment process is applied and this dataset is extended with additional information that is collected from multiple data sources. 1) Available, forsale and registered extension count related information from Extension Stats API of DOFO. 2) Extension popularity related information from Extension Information API of DOFO. 3) Language, keywords, etc. related information from DNLD API, which enabled as a result of DNLD sub-study. Then, a number of features are generated as described in detail in Chapter 4, which can be categorized into four groups: 1) sale related features such as venue (i.e. marketplace), sale weekday, sale season, etc. 2) domain name related features such as domain length, included letter count, extension, extension type, extension popularity, etc. 3) extension stats related features such as registered extension count, etc. 4) DNLD related features such as domain language, keywords, number of keywords, average keyword popularity, etc. The main focus of Chapter 4 is on preparation phase of the data for the evaluation and includes data collection, wrangling and feature extraction processes.

As illustrated in a typical machine learning flow in Figure 5.2, after feature generation in pre-processing phase, learning (prediction model training) and evaluation of the final model follow. Pre-processed data is split as train (80%) and test (20%) sets. In the learning phase, various regression models such as Linear Regression, Support Vector

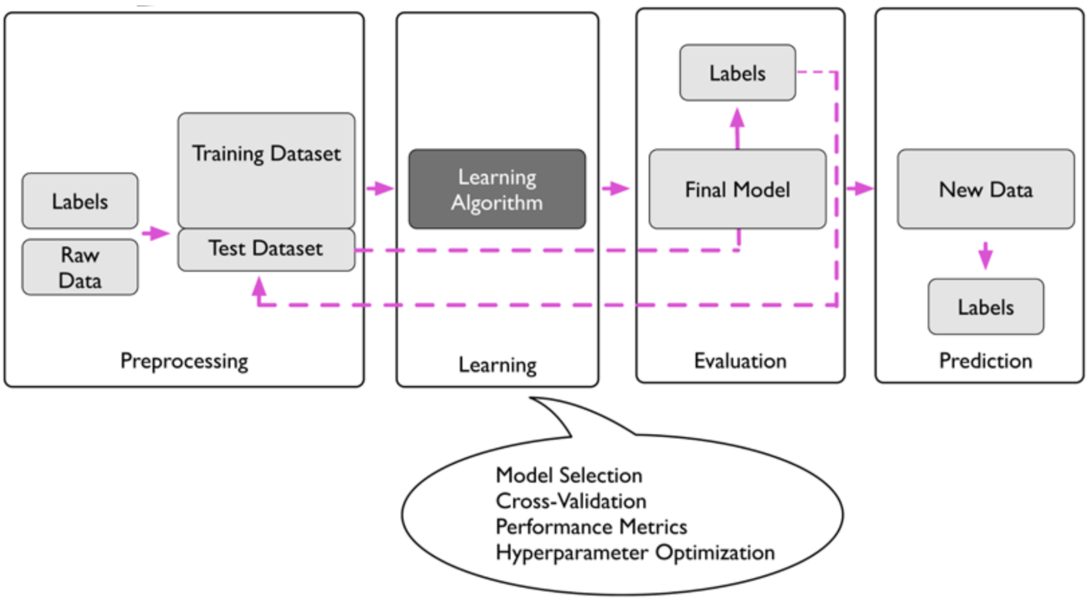


FIGURE 5.2: Machine Learning Flow [61]

Regression (SVR), Multi-Layer Perceptron Regression (MLPRegressor), Random Forest Regressor and Extreme Gradient Boosting Regression (XGBRegressor) are trained. 3-fold cross validation (CV) on train set is utilized for hyper-parameter tuning and model selection. Once final model is selected and parameters are tuned, it is evaluated with test set and the results are reported. Even-though mean absolute percentage error is used as the main evaluation criteria in this work, other regression metrics such as explained variance score, mean absolute error, mean squared error, mean squared log error, median absolute error and r2 score are also reported. All results on how well these models performed and the performance of the final model are shared in the following chapter.

Chapter 6

Results & Discussion

6.1 DNLD

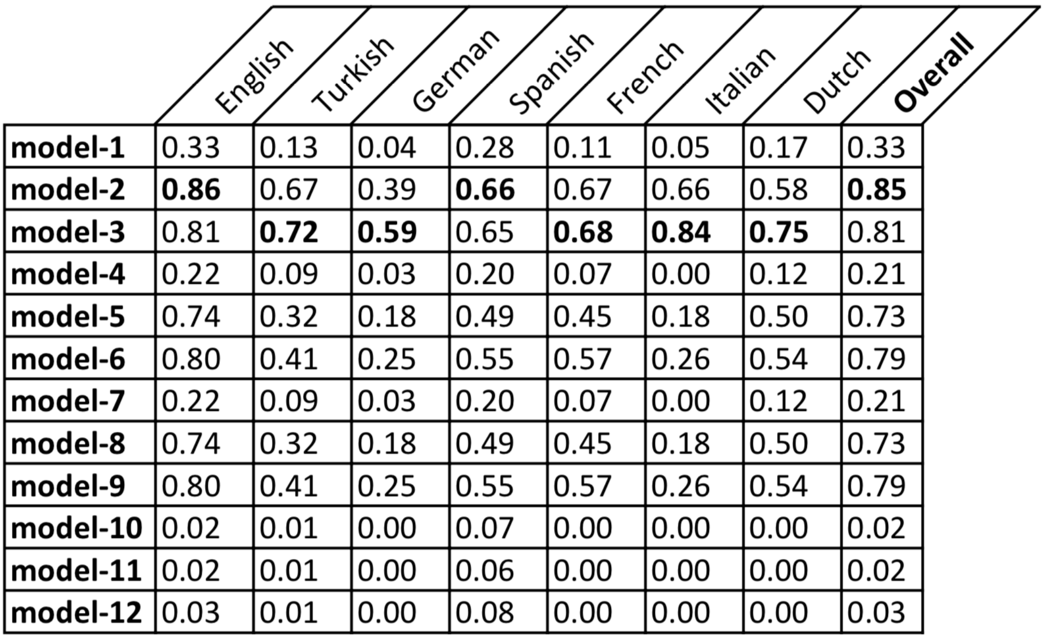
All evaluations within scope of DNLD are executed by using Noktadomains.com dataset. Description and collection process of this dataset have been provided in Section 4.1.1.2. It should be noted that it is dominated with English domain names, but majority of domain names on the Internet are English as well [62], which can also be verified with up-to-date statistics shared by W3Techs [63]. A total of 24 settings, (2 datasets: Fasttext-v1, Fasttext-v2) x (12 weights), are considered during the evaluation.

6.1.1 Keyword Extraction

Keyword extraction accuracy results are reported in Table 6.1 and Table 6.2 when DNLD is trained with Fasttext-v1 and Fasttext-v2, respectively. As can be seen in the results, there is not a single weight model and dataset combination that performs best across all languages. For example, while weight model 2 that is trained with Fasttext-v1 has performed better for English, weight model 3 that is trained with Fasttext-v2 has performed better for Turkish. It should be however noted that model 2 and model 3 consistently outperform the other weight models. Model 2 delivers the highest overall accuracy of 0.85 and 0.84 with Fasttext-v1 and Fasttext-v2, respectively. However, model 3 delivers the highest accuracy for more languages; so, it is a more constant model across languages. Overall, training with Fasttext-v1 resulted in better accuracy for English (an accuracy of 0.86 with weigh model 2), German (an accuracy of 0.59 with weigh model 2) and

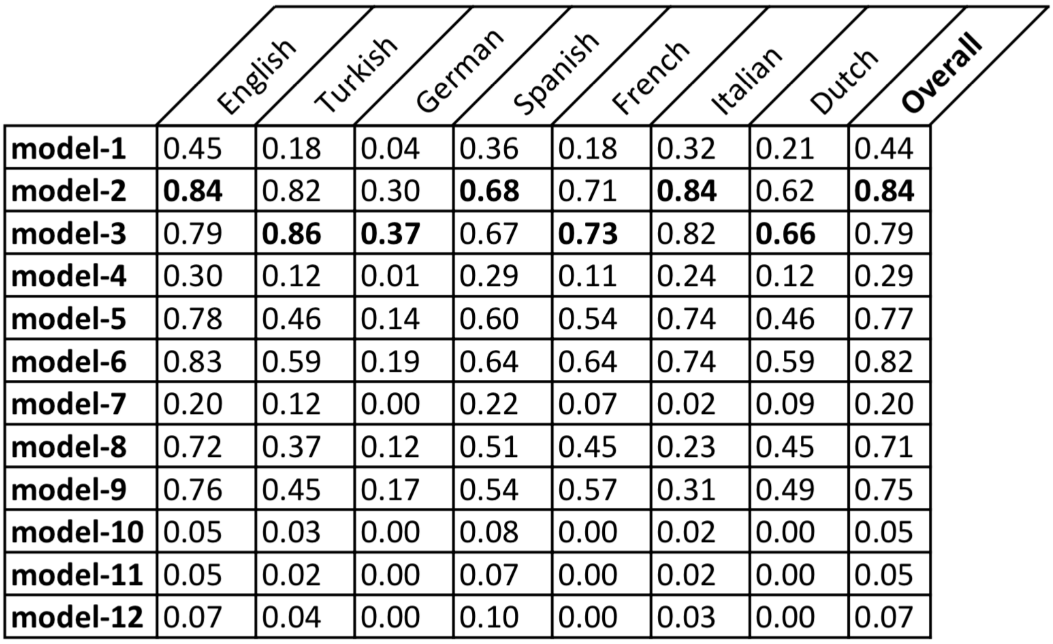
Dutch (an accuracy of 0.75 with weigh model 3) as reported in Table 6.1, and training with Fasttext-v2 resulted in better accuracy for Turkish (an accuracy of 0.86 with weigh model 3), Spanish (an accuracy of 0.68 with weigh model 2) and French (an accuracy of 0.73 with weigh model 3) as reported in Table 6.2. Keyword extraction for Italian domain names has performed equally well for both Fasttext-v1 and Fasttext-v2 with an accuracy of 0.84, but based on weigh model 3 and weight model 2, respectively.

Table 6.1: Accuracy Report for Keyword trained with Fasttext-v1



	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.33	0.13	0.04	0.28	0.11	0.05	0.17	0.33
model-2	**0.86**	0.67	0.39	**0.66**	0.67	0.66	0.58	**0.85**
model-3	0.81	**0.72**	**0.59**	0.65	**0.68**	**0.84**	**0.75**	0.81
model-4	0.22	0.09	0.03	0.20	0.07	0.00	0.12	0.21
model-5	0.74	0.32	0.18	0.49	0.45	0.18	0.50	0.73
model-6	0.80	0.41	0.25	0.55	0.57	0.26	0.54	0.79
model-7	0.22	0.09	0.03	0.20	0.07	0.00	0.12	0.21
model-8	0.74	0.32	0.18	0.49	0.45	0.18	0.50	0.73
model-9	0.80	0.41	0.25	0.55	0.57	0.26	0.54	0.79
model-10	0.02	0.01	0.00	0.07	0.00	0.00	0.00	0.02
model-11	0.02	0.01	0.00	0.06	0.00	0.00	0.00	0.02
model-12	0.03	0.01	0.00	0.08	0.00	0.00	0.00	0.03

Table 6.2: Accuracy Report for Keyword trained with Fasttext-v2



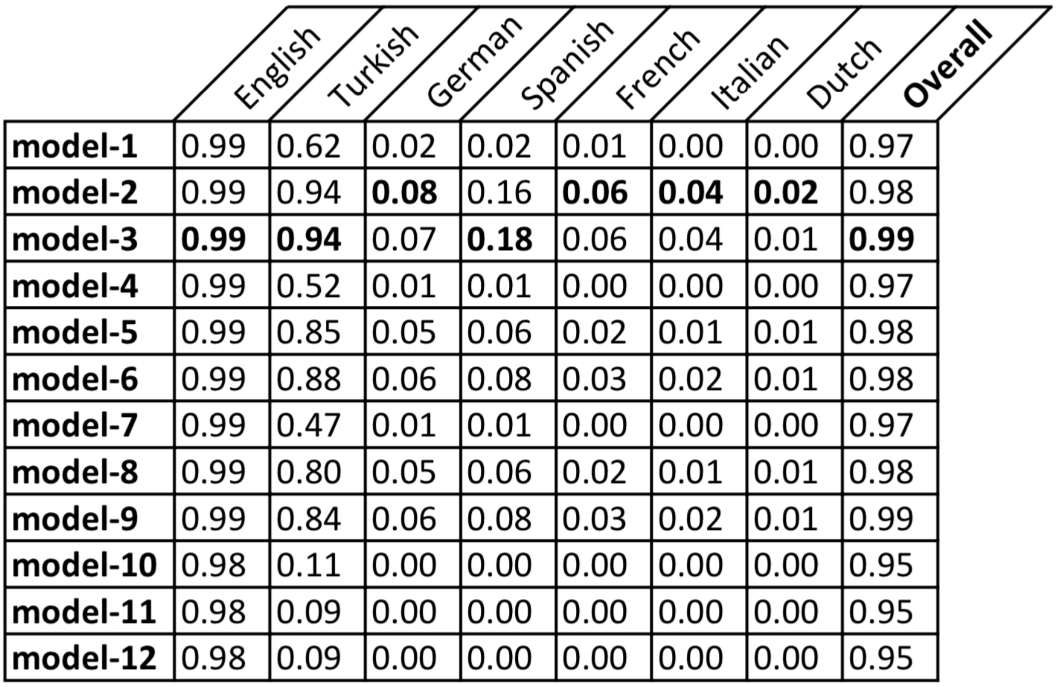
	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.45	0.18	0.04	0.36	0.18	0.32	0.21	0.44
model-2	**0.84**	0.82	0.30	**0.68**	0.71	**0.84**	0.62	**0.84**
model-3	0.79	**0.86**	**0.37**	0.67	**0.73**	0.82	**0.66**	0.79
model-4	0.30	0.12	0.01	0.29	0.11	0.24	0.12	0.29
model-5	0.78	0.46	0.14	0.60	0.54	0.74	0.46	0.77
model-6	0.83	0.59	0.19	0.64	0.64	0.74	0.59	0.82
model-7	0.20	0.12	0.00	0.22	0.07	0.02	0.09	0.20
model-8	0.72	0.37	0.12	0.51	0.45	0.23	0.45	0.71
model-9	0.76	0.45	0.17	0.54	0.57	0.31	0.49	0.75
model-10	0.05	0.03	0.00	0.08	0.00	0.02	0.00	0.05
model-11	0.05	0.02	0.00	0.07	0.00	0.02	0.00	0.05
model-12	0.07	0.04	0.00	0.10	0.00	0.03	0.00	0.07

Overall, keyword accuracy results vary between 0.59 for German and 0.86 for English. Considering the challenging nature of the problem, the results are quite encouraging. Additionally, accuracy level for each language is not directly proportional with the number of domains in dataset for the corresponding language as reported in Table 4.2. Keyword extraction accuracy for Dutch and Italian is higher than Spanish and French despite having a lower number of domain names in the evaluation dataset.

6.1.2 Language Detection

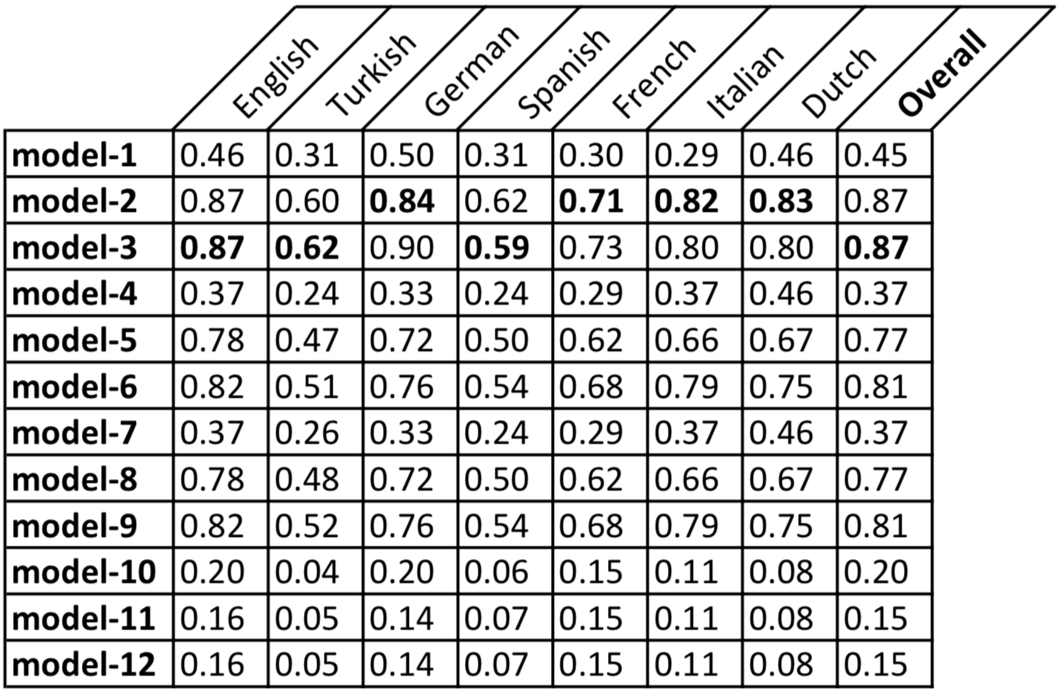
Language identification accuracy results are reported in Table 6.3, 6.4, 6.5 and Table 6.6, 6.7, 6.8 for Fasttext-v1 and Fasttext-v2, respectively. Similar to keyword extraction, there is not a single weight model and dataset combination that performs best across all languages; but model 2 and model 3 consistently outperform the others. Overall, training with Fasttext-v1 resulted in better accuracy for English (an f1-score of 0.93 with weigh model 3), Spanish (an f1-score of 0.27 with weigh model 3), French (an f1-score of 0.12 with weigh model 2), Italian (an f1-score of 0.08 with weigh model 2) and Dutch (an f1-score of 0.03 with weigh model 2) as reported in Table 6.5, and training with Fasttext-v2 resulted in better accuracy for Turkish (an f1-score of 0.84 with weigh model 3) and German (an f1-score of 0.46 with weigh model 3) as reported in Table 6.8.

Table 6.3: Precision Report for Language trained with Fasttext-v1



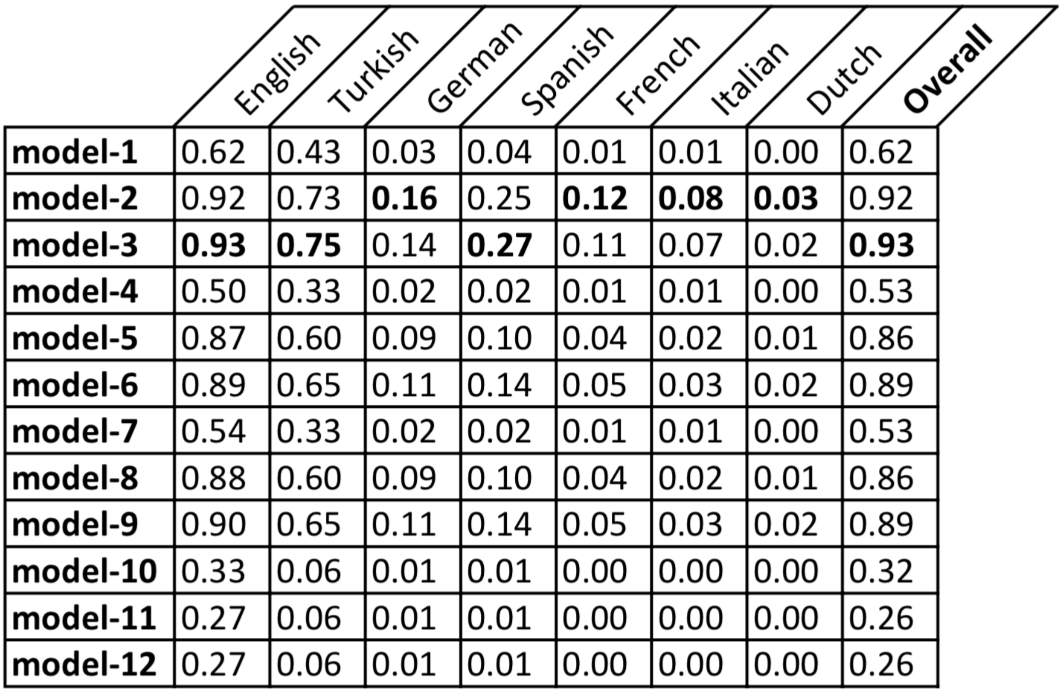
	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.99	0.62	0.02	0.02	0.01	0.00	0.00	0.97
model-2	0.99	0.94	**0.08**	0.16	**0.06**	**0.04**	**0.02**	0.98
model-3	**0.99**	**0.94**	0.07	**0.18**	0.06	0.04	0.01	**0.99**
model-4	0.99	0.52	0.01	0.01	0.00	0.00	0.00	0.97
model-5	0.99	0.85	0.05	0.06	0.02	0.01	0.01	0.98
model-6	0.99	0.88	0.06	0.08	0.03	0.02	0.01	0.98
model-7	0.99	0.47	0.01	0.01	0.00	0.00	0.00	0.97
model-8	0.99	0.80	0.05	0.06	0.02	0.01	0.01	0.98
model-9	0.99	0.84	0.06	0.08	0.03	0.02	0.01	0.99
model-10	0.98	0.11	0.00	0.00	0.00	0.00	0.00	0.95
model-11	0.98	0.09	0.00	0.00	0.00	0.00	0.00	0.95
model-12	0.98	0.09	0.00	0.00	0.00	0.00	0.00	0.95

Table 6.4: Recall Report for Language trained with Fasttext-v1



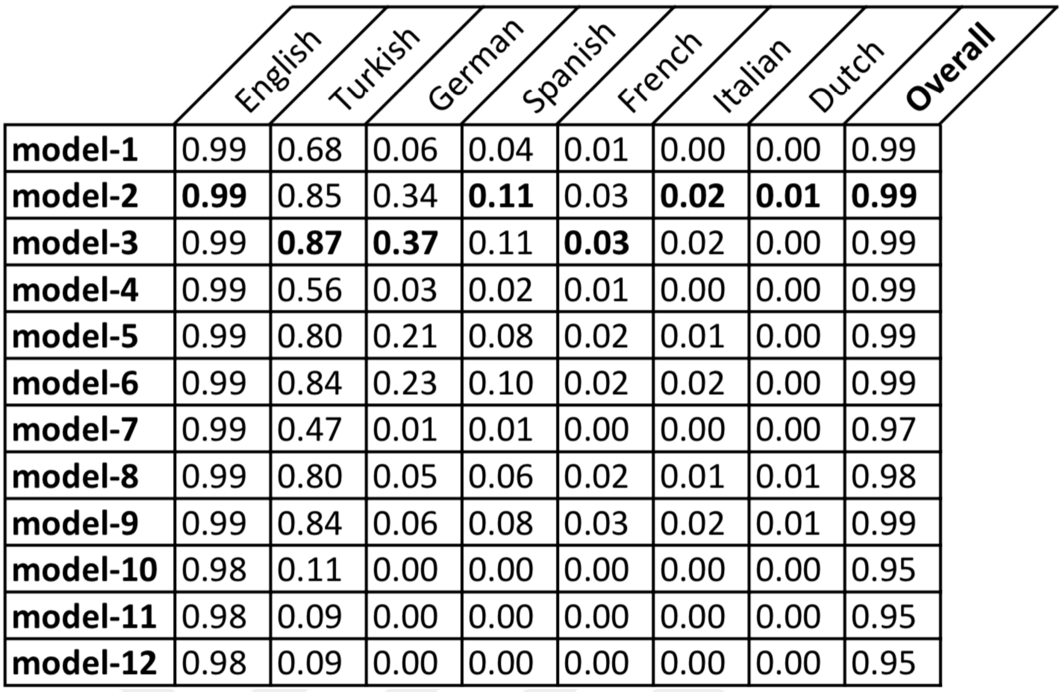
	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.46	0.31	0.50	0.31	0.30	0.29	0.46	0.45
model-2	0.87	0.60	**0.84**	0.62	**0.71**	**0.82**	**0.83**	0.87
model-3	**0.87**	**0.62**	0.90	**0.59**	0.73	0.80	0.80	**0.87**
model-4	0.37	0.24	0.33	0.24	0.29	0.37	0.46	0.37
model-5	0.78	0.47	0.72	0.50	0.62	0.66	0.67	0.77
model-6	0.82	0.51	0.76	0.54	0.68	0.79	0.75	0.81
model-7	0.37	0.26	0.33	0.24	0.29	0.37	0.46	0.37
model-8	0.78	0.48	0.72	0.50	0.62	0.66	0.67	0.77
model-9	0.82	0.52	0.76	0.54	0.68	0.79	0.75	0.81
model-10	0.20	0.04	0.20	0.06	0.15	0.11	0.08	0.20
model-11	0.16	0.05	0.14	0.07	0.15	0.11	0.08	0.15
model-12	0.16	0.05	0.14	0.07	0.15	0.11	0.08	0.15

Table 6.5: F1-Score Report for Language trained with Fasttext-v1



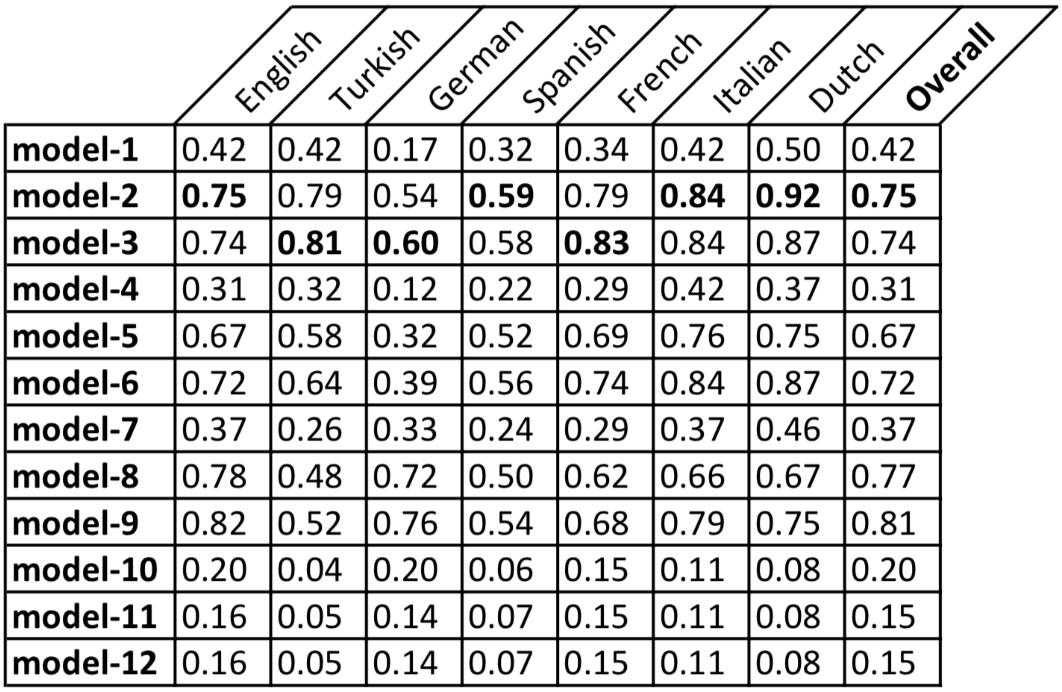
	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.62	0.43	0.03	0.04	0.01	0.01	0.00	0.62
model-2	0.92	0.73	**0.16**	0.25	**0.12**	**0.08**	**0.03**	0.92
model-3	**0.93**	**0.75**	0.14	**0.27**	0.11	0.07	0.02	**0.93**
model-4	0.50	0.33	0.02	0.02	0.01	0.01	0.00	0.53
model-5	0.87	0.60	0.09	0.10	0.04	0.02	0.01	0.86
model-6	0.89	0.65	0.11	0.14	0.05	0.03	0.02	0.89
model-7	0.54	0.33	0.02	0.02	0.01	0.01	0.00	0.53
model-8	0.88	0.60	0.09	0.10	0.04	0.02	0.01	0.86
model-9	0.90	0.65	0.11	0.14	0.05	0.03	0.02	0.89
model-10	0.33	0.06	0.01	0.01	0.00	0.00	0.00	0.32
model-11	0.27	0.06	0.01	0.01	0.00	0.00	0.00	0.26
model-12	0.27	0.06	0.01	0.01	0.00	0.00	0.00	0.26

Table 6.6: Precision Report for Language trained with Fasttext-v2



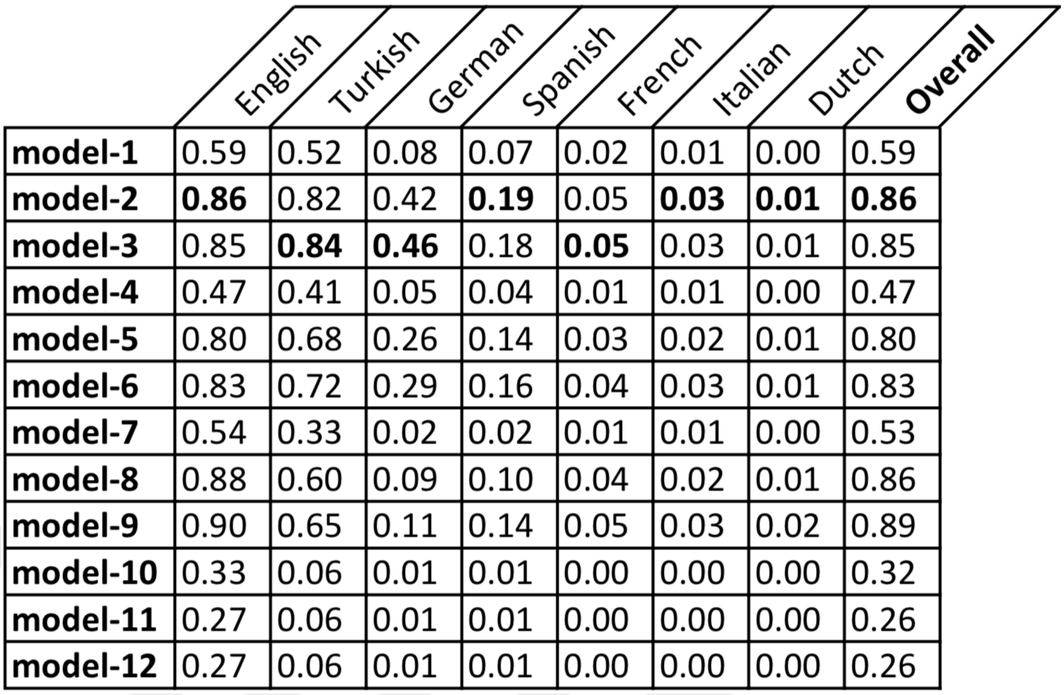
	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.99	0.68	0.06	0.04	0.01	0.00	0.00	0.99
model-2	**0.99**	0.85	0.34	**0.11**	0.03	**0.02**	**0.01**	**0.99**
model-3	0.99	**0.87**	**0.37**	0.11	**0.03**	0.02	0.00	0.99
model-4	0.99	0.56	0.03	0.02	0.01	0.00	0.00	0.99
model-5	0.99	0.80	0.21	0.08	0.02	0.01	0.00	0.99
model-6	0.99	0.84	0.23	0.10	0.02	0.02	0.00	0.99
model-7	0.99	0.47	0.01	0.01	0.00	0.00	0.00	0.97
model-8	0.99	0.80	0.05	0.06	0.02	0.01	0.01	0.98
model-9	0.99	0.84	0.06	0.08	0.03	0.02	0.01	0.99
model-10	0.98	0.11	0.00	0.00	0.00	0.00	0.00	0.95
model-11	0.98	0.09	0.00	0.00	0.00	0.00	0.00	0.95
model-12	0.98	0.09	0.00	0.00	0.00	0.00	0.00	0.95

Table 6.7: Recall Report for Language trained with Fasttext-v2



	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.42	0.42	0.17	0.32	0.34	0.42	0.50	0.42
model-2	**0.75**	0.79	0.54	**0.59**	0.79	**0.84**	**0.92**	**0.75**
model-3	0.74	**0.81**	**0.60**	0.58	**0.83**	0.84	0.87	0.74
model-4	0.31	0.32	0.12	0.22	0.29	0.42	0.37	0.31
model-5	0.67	0.58	0.32	0.52	0.69	0.76	0.75	0.67
model-6	0.72	0.64	0.39	0.56	0.74	0.84	0.87	0.72
model-7	0.37	0.26	0.33	0.24	0.29	0.37	0.46	0.37
model-8	0.78	0.48	0.72	0.50	0.62	0.66	0.67	0.77
model-9	0.82	0.52	0.76	0.54	0.68	0.79	0.75	0.81
model-10	0.20	0.04	0.20	0.06	0.15	0.11	0.08	0.20
model-11	0.16	0.05	0.14	0.07	0.15	0.11	0.08	0.15
model-12	0.16	0.05	0.14	0.07	0.15	0.11	0.08	0.15

Table 6.8: F1-Score Report for Language trained with Fasttext-v2



	English	Turkish	German	Spanish	French	Italian	Dutch	**Overall**
model-1	0.59	0.52	0.08	0.07	0.02	0.01	0.00	0.59
model-2	**0.86**	0.82	0.42	**0.19**	0.05	**0.03**	**0.01**	**0.86**
model-3	0.85	**0.84**	**0.46**	0.18	**0.05**	0.03	0.01	0.85
model-4	0.47	0.41	0.05	0.04	0.01	0.01	0.00	0.47
model-5	0.80	0.68	0.26	0.14	0.03	0.02	0.01	0.80
model-6	0.83	0.72	0.29	0.16	0.04	0.03	0.01	0.83
model-7	0.54	0.33	0.02	0.02	0.01	0.01	0.00	0.53
model-8	0.88	0.60	0.09	0.10	0.04	0.02	0.01	0.86
model-9	0.90	0.65	0.11	0.14	0.05	0.03	0.02	0.89
model-10	0.33	0.06	0.01	0.01	0.00	0.00	0.00	0.32
model-11	0.27	0.06	0.01	0.01	0.00	0.00	0.00	0.26
model-12	0.27	0.06	0.01	0.01	0.00	0.00	0.00	0.26

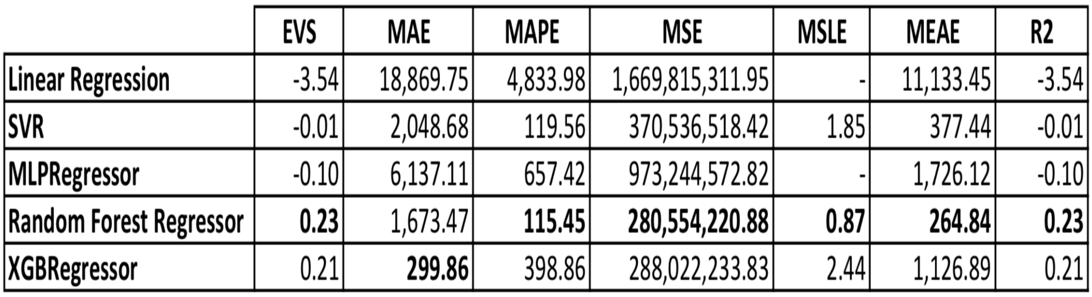
Overall performance of language detection part of DNLD varies. Although it performs quite well for English and Turkish with F1 scores 0.93 and 0.84, respectively, accuracy is rather low for French, Italian and Dutch. This can be partly explained with the unbalanced dataset we used for the reporting the results. There are also many words common across languages in Fasttext dataset; consequently having keywords in multiple languages increases the difficulty of accurate keyword extraction. Additionally, since Noktadomains.com is a Turkish company, language labeling may not be correct. However, recall scores being reasonably high across all languages is encouraging.

Performance results of Fasttext-v1 and Fasttext-v2 are close in both language detection and keyword extraction. Additionally, among the weight models, 2 and 3 perform better than the others, but they are very close to each other. However, Fasttext-v1 works slightly faster because the number of look-up words is less, overall performance of language detection with Fasttext-v1 is better and weight model 3 performed better in more languages. So, we decided to use the setup of Fasttext-v1 with weight model 3 to generate language and keyword related features when reporting DNPE results in the next section.

6.2 DNPE

All evaluations within scope of DNPE are executed by using enriched sale dataset. Data preparation and methodology have been discussed in Chapter 4 and Chapter 5, respectively. The price prediction accuracy results for various models by using a reasonable set of initial model parameters can be seen in Table 6.9. Models evaluated are Linear Regression, Support Vector Regression (SVR), Multi-Layer Perceptron Regression (MLPRegressor), Random Forest Regressor and Extreme Gradient Boosting Regression (XGBRegressor). Mean absolute percentage error is selected and reported as the main evaluation metric to compare the model performances.

Table 6.9: Performance Report by Model



	EVS	MAE	MAPE	MSE	MSLE	MEAE	R2
Linear Regression	-3.54	18,869.75	4,833.98	1,669,815,311.95	-	11,133.45	-3.54
SVR	-0.01	2,048.68	119.56	370,536,518.42	1.85	377.44	-0.01
MLPRegressor	-0.10	6,137.11	657.42	973,244,572.82	-	1,726.12	-0.10
Random Forest Regressor	**0.23**	1,673.47	**115.45**	**280,554,220.88**	**0.87**	**264.84**	**0.23**
XGBRegressor	0.21	**299.86**	398.86	288,022,233.83	2.44	1,126.89	0.21

Random Forest Regressor performed best in the initial evaluations as summarized in Table 6.9. Consequently a more extensive hyper-parameter tuning has been only performed for it due to the long run times as a result of large data size. The initial mean absolute percentage error value of 115.45 decreased to 108.84 as a result of this tuning process. Even-though mean absolute percentage error (MAPE) is utilized as main evaluation criteria to compare models, other regression metrics such as explained variance score (EVS), mean absolute error (MAE), mean squared error (MSE), mean squared log error (MSLE), median absolute error (MEAE) and r2 score (R2) are also reported.

Table 6.10: Tuned Performance Report for Random Forest Regressor

Regression Metric	**Value**
explained variance score	0.22
mean absolute error	1,642.82
mean absolute percentage error	**108.84**
mean squared error	288,193,615.49
mean squared log error	0.77
median absolute error	244.10
r2 score	0.22

6.2.1 Eliminating Outliers

Loosely speaking, an outlier is a data point that has a significant 'distance' to other observations. Detection and elimination of outliers can be crucial in ML model training as they can substantially deteriorate model performance and they are also usually not the point of interest in prediction as well. Figure 4.3 indicates that the domain sale price distribution is highly right skewed. Moreover, the gap between mean (2,640.935) and median (500) is quite large as seen in Table 4.4. These are strong indicators of outliers. A close look at the data shows that there are extreme outliers such as a domain sale for 14,000,000. We do not expect the price of these extreme outliers to follow a regular pattern that can be learned by ML algorithms and this type of sales is not typically done through regular sale channels; so they are not the target of this study. Consequently, we explore outliers elimination methods and study the effect on the accuracy. 3 different approaches and the corresponding thresholds from loose to strict limits are explored.

6.2.1.1 Standard Deviation Method (STD)

Standard deviation method is generally preferred in cases where the distribution is Gaussian-like. 3 standard deviation (std) is one of the common metrics to set the threshold. Standard deviation is quite large for our dataset with extreme outliers and lead to a very loose threshold. The formula is (mean + 3 * std), so $2,640.9 + 3 * 51,224.7$ equals 156,315. This also corresponds to e=12 in Figure 4.3 that is on the extreme right tail of the distribution and a range that could be safely excluded in our study.

Table 6.11: Domain Sale Price Stats with 3 STD

	price
count	586,734
mean	1,836.9
std	5,850.0
min	90
25%	200
50%	500
75%	1,700
max	155,688

Although the gap between mean (1,836.9) and median (500) is quite large as reported in Table 6.11, it is decreased by ~804 after 3-std filter applied to the original dataset. Prediction results in Table 6.12 show substantial improvement, mean absolute percentage error is going down to 80.17% from 108.84%.

Table 6.12: Tuned Performance Report for Random Forest Regressor with 3 STD

Regression Metric	**Value**
explained variance score	0.46
mean absolute error	1,085.72
mean absolute percentage error	**80.17**
mean squared error	18,645,410.60
mean squared log error	0.63
median absolute error	219.67
r2 score	0.45

6.2.1.2 Center 99%

We considered center 99% of the data and eliminated the remaining 1% as outliers that resulted in a threshold of 24,500.

Table 6.13: Domain Sale Price Stats with 99% Percentile

	price
count	581,486
mean	1,406.7
std	2,455.9
min	90
25%	200
50%	500
75%	1,638
max	24,485

It should be noted that the gap between mean (1,406.7) and median (500) is decreased by ~1,234 when outlier elimination based on this threshold is applied to the original dataset as reported in Table 6.13. Domain name price prediction accuracy was further improved as reported in in Table 6.14. Mean absolute percentage error was reduced to 72.42%.

Table 6.14: Tuned Performance Report for Random Forest Regressor with 99% Percentile

Regression Metric	**Value**
explained variance score	0.45
mean absolute error	734.57
mean absolute percentage error	**72.42**
mean squared error	3,175,930.43
mean squared log error	0.58
median absolute error	212.75
r2 score	0.44

6.2.1.3 Interquartile Range Method (IQR)

Interquartile range method for outlier identification is generally preferred in cases where the distribution is non-Gaussian. It is calculated based on the difference between 75th and the 25th percentiles of the data. To get the upper bound, this difference is multiplied with 1.5 and summed with 75th. The formula is ((75th - 25th) * 1.5 + 75th), so $(1,710-200)*1.5+1,710$ equals 3,975.

Table 6.15: Domain Sale Price Stats with IQR

	price
count	536,741
mean	855.3
std	914.7
min	90
25%	187
50%	410
75%	1,250
max	3,975

The gap between mean (855.3) and median (410) is decreased by ~1,785 after IQR filter applied to the original dataset as reported in Table 6.15. A substantial jump in domain name price prediction accuracy is observed as reported in in Table 6.16. Mean absolute percentage error was improved to 58.54%.

Table 6.16: Tuned Performance Report for Random Forest Regressor with IQR

Regression Metric	**Value**
explained variance score	0.60
mean absolute error	352.99
mean absolute percentage error	**58.54**
mean squared error	347,620.74
mean squared log error	0.45
median absolute error	166.00
r2 score	0.59

At the final stage, the upper bound for the elimination of outliers is set to 3,975 with the help of IQR. Generally, for domain names that are more valuable than a few thousand dollars, the human factor and consequently an increasing uncertainty is starting to come into play. For this reason, it is possible to say that, this threshold actually corresponds to the desired range, where a regular pattern in pricing structure can be detected by

ML models. This is also the range that an automated accurate prediction has a greater business value. As seen in results, the prediction accuracy consistently increases as more strict outlier elimination criteria is applied. A very encouraging and acceptable mean absolute percentage error of 58.54 is achieved after IQR based outlier elimination.

Chapter 7

Conclusion

Within the scope of this thesis, the question of whether it is possible to predict the value of a domain name is examined with a Machine Learning (ML) based approach and DNPE is proposed as a result. An extensive domain name sale history dataset is collected as part of this study and numerous unique features are extracted based on domain name. Identification of domain language and the extraction of words within a domain name is essential in representing the domain name characteristics that have profound effect on its value such as the number of words and the popularity of words used in domain name. So, DNLD, which can support up to 265 languages depending on Fasttext datasets, is proposed in order to address this need. Additionally, it is modeled as a weighted interval scheduling problem and a number of weighting schemes are proposed as well.

Keyword accuracy results of DNLD vary between 0.59 for German and 0.86 for English. Moreover, language accuracy results are quite well for English and Turkish with F1 scores 0.93 and 0.84, respectively. The possible reasons why other languages did not perform as good as these two, such as unbalanced and inaccurate dataset and the appearance of same keywords in multiple languages, have been discussed. And for the DNPE, we highlighted that outliers have a substantial effect on the accuracy and a mean absolute percentage error of 58.54 has been achieved after outlier elimination with IQR. As a general assessment, the results for both DNLD and DNPE parts of the study are quite encouraging, considering the complexity of the problem and uncertainty in domain name prices. As a pioneering work in domain name language detection, keyword extraction and price prediction, we believe the approach and methodologies developed in this thesis would pave the way for subsequent studies and further advancements in future.

Chapter 8

Future Work

Weight model definitions evaluated as part of DNLD sub-study are pre-defined. As a future work, parameterizing the weight definition as in the following formula as an example enables the consideration of a larger set of weight definitions by varying the variables a, b and c. In this approach, variables could be treated as algorithm hyper-parameters and tuned. Or it may even be possible to utilize parameterized weight definition as model and learn the model parameters with an ML approach. A more extensive search of the best weight definition can help improving the accuracy of DNLD.

$$a * word\_length^b * (\frac{word\_rank}{word\_count})^c$$

The value of a domain name depends on its properties. In this study, we did our best to characterize many important properties of a domain name including the ones highlighted by others as reviewed in Chapter 2 and transformed them into features as much as possible. However, as can be expected, some properties were not modeled due to the lack of data or methodology. We believe properties such as memorability and readability [52] can have a profound effect on the accuracy of domain name prediction. We plan to explore methods to model these properties and study their effect on the price prediction in a future work.

Deep Neural Network (DNN) based language models have gained popularity in recent years and used in many NLP tasks. A DNN-based language model, particularly trained at character-level, can be an excellent fit for domain name language identification and keyword extraction. We plan to explore the use of language models on these tasks in future.

Bibliography

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[2] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*, 2018.

[3] A. Maurushat. *Ethical Hacking.* University of Ottawa Press, 2019.

[4] Domain name structure, . https://firstsiteguide.com/choose-domain/.

[5] Unicode. http://www.unicode.org.

[6] A. Costello. Punycode: A bootstring encoding of unicode for internationalized domain names in applications (idna). Technical report, 2003.

[7] Icann. https://www.icann.org/resources/pages/rfcs-2012-02-25-en.

[8] Dofo blog. https://blog.dofo.com/internationalized-domain-names/.

[9] W. Hanson. Principles of internet marketing, cincinnati. *Ohio: South-Western College Publishing*, 2000.

[10] The domain name industry brief, 2019. https://www.verisign.com/assets/domain-name-report-Q12019.pdf.

[11] J. Levine. How big is the domain business? http://www.circleid.com/posts/20180813_how_big_is_the_domain_business/.

[12] J. Zoch. Uniregistry brokerage boasts impressive 2016-2018 sales. https://uniregistry.com/blog/post/uniregistry-brokerage-boasts-impressive-2016-2018-sales.

[13] Govalue. https://www.godaddy.com/domain-value-appraisal.

[14] Estibot. https://www.estibot.com.

[15] Epik. https://appraise.epik.com.

[16] Freevaluator. https://www.freevaluator.com.

[17] Websiteoutlook. https://www.websiteoutlook.com.

[18] E. Baykan, M. Henzinger, and I. Weber. Web page language identification based on urls. *Proceedings of the VLDB Endowment*, 1(1):176–187, 2008.

[19] E. Baykan, M. Henzinger, and I. Weber. A comprehensive study of techniques for url-based web page language classification. *ACM Transactions on the Web (TWEB)*, 7(1):3, 2013.

[20] T. A. Abdallah and B. de La Iglesia. Url-based web page classification: With n-gram language models. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 19–33. Springer, 2014.

[21] M. Kan. Web page classification without the web page. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 262–263. ACM, 2004.

[22] E. Baykan, M. Henzinger, L. Marian, and I. Weber. Purely url-based topic classification. In *Proceedings of the 18th international conference on World wide web*, pages 1109–1110. ACM, 2009.

[23] B. Martins and M. J. Silva. Language identification in web pages. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 764–768. ACM, 2005.

[24] V. Vega. Continuous-learning weighted-trigram approach for indonesian language distinction: A preliminary study. In *In Proceedings of 19th International Conference on Computer Processing of Oriental Languages.(2001) Vinsensius Berlian Vega SN and Stéphane Bressan.* Citeseer, 2001.

[25] K. Somboonviwat, M. Kitsuregawa, and T. Tamura. Simulation study of language specific web crawling. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1254–1254. IEEE, 2005.

[26] N. C. Ingle. A language identification table. *The Incorporated Linguist*, 15(4), 1976.

[27] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67, 1999.

[28] G. Grefenstette. Comparing two language identification schemes. In *Proceedings of JADT*, volume 95, 1995.

[29] T. Dunning. *Statistical identification of language.* Computing Research Laboratory, New Mexico State University Las Cruces, NM, USA, 1994.

[30] W. J. Teahan and D. J. Harper. Using compression-based language models for text categorization. In *Language modeling for information retrieval*, pages 141–165. Springer, 2003.

[31] E. Tromp and M. Pechenizkiy. Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34, 2011.

[32] R. Řehřek and M. Kolkus. Language identification on the web: Extending the dictionary method. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 357–368. Springer, 2009.

[33] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer, 1994.

[34] K. Hayati. Language identification on the world wide web. *Masters project, University of California, Santa Cruz*, 2004.

[35] P. Sibun and J. C. Reynar. Language identification: Examining the issues. 1996.

[36] P. Norvig. Natural language corpus data. *Beautiful data*, pages 219–242, 2009.

[37] Z. Bikadi, S. Ahangama, and E. Hazai. Prediction of domain values: High throughput screening of domain names using support vector machines. *arXiv preprint arXiv:1707.00906*, 2017.

[38] S. Dieterle and R. Bergmann. Case-based appraisal of internet domains. In *International Conference on Case-Based Reasoning*, pages 47–61. Springer, 2012.

[39] S. Dieterle and R. Bergmann. A hybrid cbr-ann approach to the appraisal of internet domain names. In *International Conference on Case-Based Reasoning*, pages 95–109. Springer, 2014.

[40] A. Tajirian. Statistical models for market approach to domain name valuation. *DomainMart. com*, 2010.

[41] J. H. Tang, M. C. Hsu, T. Y. Hu, and H. H. Huang. A general domain name appraisal model. *Journal of Internet Technology*, 15(3):427–431, 2014.

[42] R. M. Visconti. Domain name valuation: Internet traffic monetization and it portfolio bundling. 2017.

[43] Z. Wu and H. He. Domain name valuation model constructing and emperical evidence. In *2009 International Conference on Multimedia Information Networking and Security*, volume 2, pages 201–204. IEEE, 2009.

[44] Z. Wu, G. Zhu, R. Huang, and B. Xia. Domain name valuation model based on semantic theory and content analysis. In *2009 Asia-Pacific Conference on Information Processing*, volume 2, pages 237–240. IEEE, 2009.

[45] J. R. Gould and S. J. Coyle. How consumers generate clickstreams through web sites: An empirical investigation of hypertext, schema and mapping theoretical explanations. *Journal of Interactive Advertising*, 2(2):42–56, 2002.

[46] T. Lindenthal. Valuable words: The price dynamics of internet domain names. *Journal of the Association for Information Science and Technology*, 65(5):869–881, 2014.

[47] T. Lindenthal. Monocentric cyberspace: The primary market for internet domain names. *The Journal of Real Estate Finance and Economics*, 57(1):152–166, 2018.

[48] M. L. Mueller. The battle over internet domain names: Global or national tlds? *Telecommunications Policy*, 22(2):89–107, 1998.

[49] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. Domain bias in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 413–422. ACM, 2012.

[50] J. S. Ilfeld and R. S. Winer. Generating website traffic. *Journal of Advertising Research*, 42(5):49–61, 2002.

[51] J. Murphy, L. Raffa, and R. Mizerski. The use of domain names in e-branding by the world's top brands. *Electronic Markets*, 13(3):222–232, 2003.

[52] Domaintools, . `https://www.domaintools.com/support/domain-valuation-how-do-i-value-a-domain-name`.

[53] The global domain name market in 2017, 2018. `https://www.afnic.fr/medias/documents/etudes/Global_Domain_Name_Market_in_2017_FINAL.pdf`.

[54] Dofo. `https://dofo.com`.

[55] Fasttext. `https://fasttext.cc`.

[56] Noktadomains. `https://www.noktadomains.com`.

[57] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

[58] K. Thein. Apache kafka: Next generation distributed messaging system. *International Journal of Scientific Engineering and Technology Research*, 3(47):9478–9483, 2014.

[59] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja. Lambda architecture for cost-effective batch and speed big data processing. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2785–2792. IEEE, 2015.

[60] A. W. Kolen, J. K. Lenstra, C. H. Papadimitriou, and F. C. Spieksma. Interval scheduling: A survey. *Naval Research Logistics (NRL)*, 54(5):530–543, 2007.

[61] S. Raschka and V. Mirjalili. *Python machine learning*. Packt Publishing Ltd, 2 edition, 2017.

[62] G. Grefenstette and J. Nioche. Estimation of english and non-english language use on the www. In *Content-Based Multimedia Information Access-Volume 1*, pages 237–246. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

[63] W3techs. `https://w3techs.com/technologies/history_overview/content_language`.