MEF UNIVERSITY

DUPLICATE RECORD DETECTION: A RULE-BASED APPROACH

Capstone Project

Gülce Malkaralı

İSTANBUL, 2017

MEF UNIVERSITY

DUPLICATE RECORD DETECTION: A RULE-BASED APPROACH

Capstone Project

Gülce Malkaralı

Advisor: Prof. Dr. Özgür Özlük

İSTANBUL, 2017

EXECUTIVE SUMMARY

DUPLICATE RECORD DETECTION:
A RULE-BASED APPROACH

Gülce Malkaralı

Advisor: Prof. Dr. Özgür Özlük

SEPTEMBER, 2017, 18 Pages

This study presents a rule-based algorithm to detect duplicate and near-duplicate records within a dataset that is extracted from a leading online realty platform. The aim is to improve end-user experience through reducing the volume of repetitive content and provide more accurate and varied search outcomes. The rule-based algorithm provided the opportunity to detect similarities under sub-categories and different status groups with superior outcomes in comparison to the existing software. It also demonstrated measurable outcome performance and definition of similarities. Further steps to this study could be diversification and redefinition of status groups through machine learning classifiers.

Key Words: Duplicate Record Detection, Near Duplicate, Similarity Function, Spam Score.

ÖZET

MÜKERRER İLAN TESPİTİ:
KURALA DAYALI BİR YAKLAŞIM

Gülce Malkaralı

Tez Danışmanı: Prof. Dr. Özgür Özlük

EYLÜL, 2017, 18 Sayfa

Bu çalışma, lider çevrimiçi emlak platformlarından biri tarafından sağlanan veri seti üzerinde mükerrer ve neredeyse mükerrer olan girdilerin tespit edilmesine yönelik kural bazlı bir algoritma geliştirilmesini ele almaktadır. Bu sayede, mükerrer içeriğin azaltılarak daha doğru ve çeşitli sonuçların öne çıkarılması; dolayısıyla son kullanıcı deneyiminin iyileştirilmesi amaçlanmaktadır. Söz konusu kurala dayalı algoritma, alt kategoriler bazında ve farklı statü gruplarına göre benzerlikleri tespit edebilmekte ve mevcut olan versiyona göre daha iyi sonuçlar vermektedir. Ayrıca gerek mükerrer olma durumunun tanımlanması gerek benzerlik oranlarının tespiti çok daha ölçülebilir hale getirilmiştir. Çalışmanın daha ileri adımlarında alt grupların çeşitlendirilmesi ve mükerrerlik koşulunun yeniden tanımlanabilmesi için makine öğrenmesi sınıflandırıcılarından destek alınması değerlendirilebilir.

Anahtar Kelimeler: Mükerrer İlan Tespiti, Benzerlik Fonksiyonu, Spam Puanı.

TABLE OF CONTENTS

1. INTRODUCTION

The study examines the detection of duplicate and near-duplicate records through a rule-based approach. The aim of the project is to improve user experience quality by reducing the number of entities addressing the same content. For this purpose, a rule based algorithm is developed and later fine-tuned in line with the business objectives at hand. This algorithm aims to provide a reliable solution to identify the genuine and repetitive records based on the selected features and business priorities. In this chapter, the objectives and scope of the study will be dwelled upon in detail along with the brief literature review that highlights significance of the issue.

1.1. Duplicate Record Detection: A Brief Literature Survey

User generated content on the popular platforms has been proliferated in the recent years as a result of scalable digital transformation of services. Consequently, similar record detection became a significant field of study to 'reduce the data volume and increase the search efficiency' (Lin et al., 2013, p. 1467). Several methods were developed to deal with duplicate record problem that varies in line with how the duplication is actually defined. In this section, a couple of examples will be reviewed as a form of brief literature survey.

The similarity problem covers two closely related yet different categories: duplicates and near duplicates (Lin et. al, 2013). If modification exists in terms of insertion, deletion or replacement to a predetermined degree, issue of near-duplicate records prevails which is evident in almost all kinds of social media (Lin et. al, 2013). The fundamental question becomes to a what extent/degree the similarities between the distinct entities will be tolerated or alerted. In that regards, the predetermination of near-duplicate records builds on two important pillars: feature/label selection or feature/label formation. Since most of the social media or website entities are in free text format, several studies detect duplications by first utilizing bag of words method and then applying classifiers. In "Uncovering Social Spammers: Social Honeypots and Machine Learning", Lee et. al. aimed to predict future spam behavior by 'developing machine learning classifiers for identifying previously unknown spammers with high precision' (Lee et al., 2010, p.1) To achieve that, they utilized bag of words and then applied SMV classifiers. This approach was repeated multiple times with additions that focus on feature vectors and alternating

decision trees (Martin, 2011). Recently, Quora opened up its dataset on Kaggle platform to revolutionize their random forest based classification approaches with neural networks.[1]

1.2. About Hürriyet Emlak

The dataset for this study is graciously provided by Hürriyet Emlak within the boundaries of a non-disclosure agreement.[2] The unmasked dataset was handed over for academic purposes only and Hürriyet Emlak intervened neither the methods nor the scope of the study. Two onsite meetings and an online session were held in order to support and guide the project.

Hürriyet Emlak, founded in 2006, is the most popular real estate focused online platform operating in Turkey[3]. The platform, which reached the volume of one million daily active advertisements, brings together realty firms, individual members and end-users under through innovative services. The goals of the platform consist of being the first choice among the realty firms, reducing the burden of realty search thanks to correct and updated advertisements and providing new services[4]. In line with the mission statements of the platform, detection of duplicate records was given great importance due to its potential in increasing user experience quality.

1.3. About the Dataset

The dataset shared by Hürriyet Emlak consisted of all the reality advertisements from Istanbul's Ümraniye district for a given period. The total number of rows, or the number of unique realties in other words, were 4970 in the beginning. However, the number of rows/instances were reduced from 4970 to 3137 prior to the data analysis. It was noticed that there were repetitions caused by the data mining process. In other words, these were not duplicates that were intended to be detected within the scope of this project. Rather than that, they were caused by technicalities only, i.e. each time an advertiser

---

[1] Kaggle. Quora Question Pairs. Retrieved from: https://www.kaggle.com/c/quora-question-pairs

[2] The non-disclosure agreement was signed on 07.04.2017 at Hürriyet Emlak headquarters.

[3] Hürriyet Emlak Corporate Identity. Retrieved from: http://www.hurriyetemlak.com/Main/pgAboutUs.aspx

[4] For instance, Hürriyet Emlak 'Temiz İlan Hattı' is a hotline for realties that had been let or sold despite active advertisements. Retrieved from: https://www.hurriyetemlak.com/temiz-ilan-donemi-basliyor/emlak-yasam-sektorden-haberler/96n3N2v9V20=

logged in their account (in spite of any updates) the unique entry was multiplied which did not signify any statistically important and rationally meaningful relationship. This reduction in row number from 4970 to 3137 stands for as the first step of data preprocessing albeit a primitive one. All the model building and algorithm development is based on the version that involves 3137 rows in total.

The dataset involves all the possible information available for each realty advertisement. It includes total of 104 columns or potential features to be used for the analysis. Examples for these features are price, advertisement type (rental or sale), realty type (house/residential, workplace, land), neighborhood and district, building age, room number, square meter, floor number, heating type etc. These features predominantly have integer values either in the form of continuous numerical values (price, square meter, age, etc.) or category ID's (coding for building type, heating type, currency type etc.). The only exceptions to these are Realty Title and Realty Description which stand for free text format. Consequently, the dataset seems suitable for classification algorithms considering the abundance of features/labels.

In addition to the instance and feature numbers, it should also be noted that all of the advertisements are generated by 194 distinct realty firms or individual members. The portfolio distribution among those varies unevenly between 2 and 301 realty records per Firm ID. As it will be explained in detail shortly, the realty firm information became an integral part of the study.

Last but not least, the dataset provided by Hürriyet Emlak also includes another sheet which presents the previous algorithm's outcomes. The file, smiler.csv, lists the detected duplicate records as pairs. It is consisted of 2486 rows and 2 columns (Realty ID and Duplicate Realty ID). The outcomes of this file did not set as the performance target to be met or surpassed. Rather than that, it was presented as a reference document since the basis of the detection or presentation of the outcomes was unknown (closed software). Consequently, clarification of the existing algorithm and measuring its performance became another focal point of the study.

2. PROJECT STATEMENT AND METHOLODY

In this section, the objective and scope of the project will be discussed by highlighting the business priorities. Following that, the methodology will be presented covering steps such as exploratory data analysis and model deployment.

2.1. Problem Statement

The problem statement is refined during the meetings held with Hürriyet Emlak. The duplicate and near-duplicate record detection is particularly important to prevent the advertisers to enter their portfolios to the portal multiple times despite any updates. Those repetitions can either be intentional or unintentional, so the proliferation of duplication is not necessarily defined as spamming (the ratio of all records to duplicate records for each firm is measured though). In other words, the similarities found in the dataset are needed to be alerted only when they were entered by the same realty office/individual member. Therefore, duplicate detection needs to be controlled under a sub-category; that is 'Firm ID'. Consequently, the duplicate record detection algorithm should be run in each distinct 'Firm ID' (total of 194).

To sum up, in line with the business priorities, the definition of the duplication and near-duplication is determined to be valid only under same Firm ID. That paved the way for decision to proceed with rule-based algorithm rather than machine learning classifiers which will be explained in detail shortly.

2.1.1 Project Objectives

The objective of the project is to develop a simple yet precise algorithm that detects duplicate and near-duplicate records. In the meantime, understanding how the existing similarity detection algorithm works and measuring its performance became another fundamental objective.

The breakdown of project objectives are as follows:

- Analyzing the dataset provided by Hürriyet Emlak
- Developing a duplicate detection algorithm (including feature selection process) that is based on realty firms
- Processing the data and alerting when the duplicate or near-duplicate record(s) was found (demonstrating the realty IDs)

- Measuring performance of the algorithm and the existing version (comparison of outcomes)

2.1.2 Project Scope

The scope of the process covers developing an original, simple and precise detection algorithm and understanding/measuring the existing algorithm. In line with the business side's brief, five different status groups are defined which signify for different levels/definitions of similarity. The first status group stood for direct duplication while the other four demonstrated distinct conditions of near-duplicate situation. First, total number of duplication and near-duplication cases are found. Then these duplications are divided under different status groups. The Python algorithm presented the opportunity to extract several csv files for business purposes. In the meantime, the outcome of the existing algorithm is modified according to the business priorities. The duplications found by the smiler.csv file was not based on the Firm ID subcategories so it was needed to be transformed in line with the problem statement and project objectives.

The definition of these five different status groups are built on business insight, intuition and exploratory data analysis (mainly missing value analysis). As a next step, statistical significance of them (R-Square) of this model is calculated. However, the fine-tuning of sub-category boundaries (such as minor sale price or square meter alterations that defines near-duplication) are not falling in the scope of this study. Since the rule-based approach would be always effective when the computational values are changed, that decision (values of boundaries) rests on the business side.

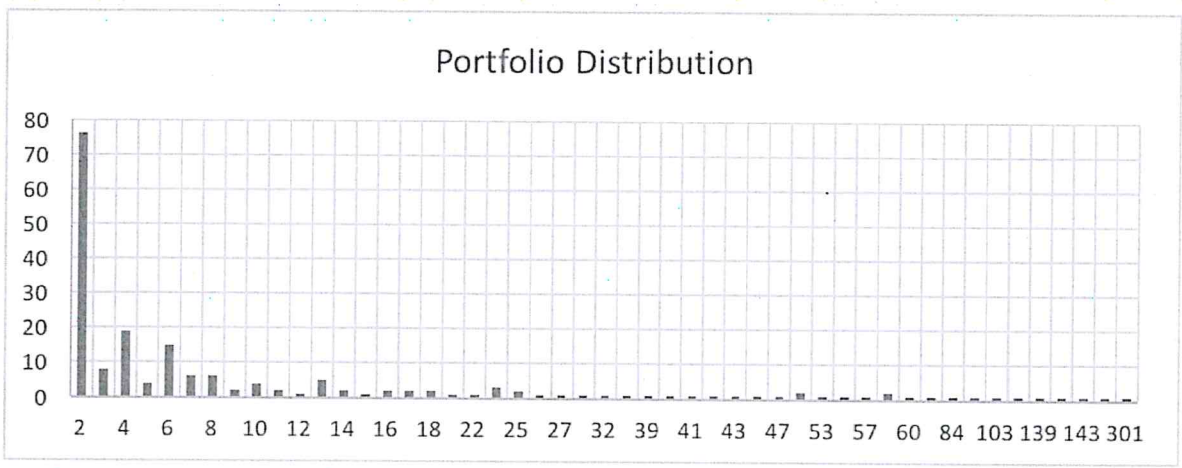
2.2. Methodology

In line with the problem statement and project objectives, a rule-based approach rather than development of a machine learning algorithm is preferred. The rule-based algorithm targets ultimate precision in detection of duplicate and near-duplicate records which would not possible with a machine learning algorithm (at least according to the preferred standards).

The dataset is composed of 3137 unique rows, property advertisements, and 194 unique realty firm ID's. The distribution of the property advertisements (3137 rows) among the firms are quite unbalanced. Only 58 out of 194 firms/individual members have more than 10 properties on their portfolio (in a given district) while 78 of them have just 2

properties. On the contrary, 7 firms have more than 100 properties, one being the most distinct outlier with 301 properties.[5] Consequently, since the distinct number of firms are high along with the respective unbalanced distribution, applying machine learning classifiers under each subcategory, i.e. Firm ID, would not be lead to significant outcomes.

Figure 1: Distribution of Property Portfolio among the Firms



The rule-based algorithm is developed in Python interpreter (Anaconda Spyder) by mostly utilizing Numpy and Pandas libraries. Since the volume of the dataset and the objectives and scope of the project did not enable using feature selection methods, the features that are used for model building were selected based on business insight and exploratory data analysis.

2.2.1 Exploratory Data Analysis

During the exploratory data analysis phase, the impact of features on the sale price was measured through correlation tests and linear regression model. Although price prediction does not constitute the focal point of this study, the adjusted R-square scores became important indicators for the impact of the features.

In addition to data visualization and analysis of correlation and adjusted R-square scores, the ratio of missing values in each feature is analyzed as well. It has been demonstrated that although the dataset is rich in terms of labels (104 different features for each unique row), 37 of these features are completely consisted of missing values.

---

[5] The statistical summary of portfolio distribution among the Firm ID's are as follows: Min: 2, 1st Qu.: 2, Median: 4, Mean: 16,17, 3rd Qu.: 13, Max: 301.

Furthermore, 50 of these 104 features have more than 70% missing values making them inconvenient for model building considering the limited volume of the dataset.

Table 1: Features' Missing Value Ratio Analysis

Features No.	NA Ratio
39	<20
42	<30
43	<40
45	<50

The main motivation for this phase is to simplify the model by reducing the number of features by selecting the ones that are the most significant. To do so, the features with the least missing values are targeted. As a matter of fact, the significance and the minimum ratio of missing values are de-facto correlated (the advertiser needs to fill out certain information/bare minimum to successfully generate an entity). Since the precision of the rule-based approach is important, imputing the missing values was not preferred. The algorithm could still catch similarities when the same features of both entities have missing values. That being said, selecting the features with the great number of missing values would not be advantageous to detect the similarities. Consequently, features with low ratio of missing values are included by leaving the missing values as-is in the model with the exception of Floor ID. Despite this feature's relatively higher ratio of NA values, it is significant especially for one of the status groups which will be explained shortly.

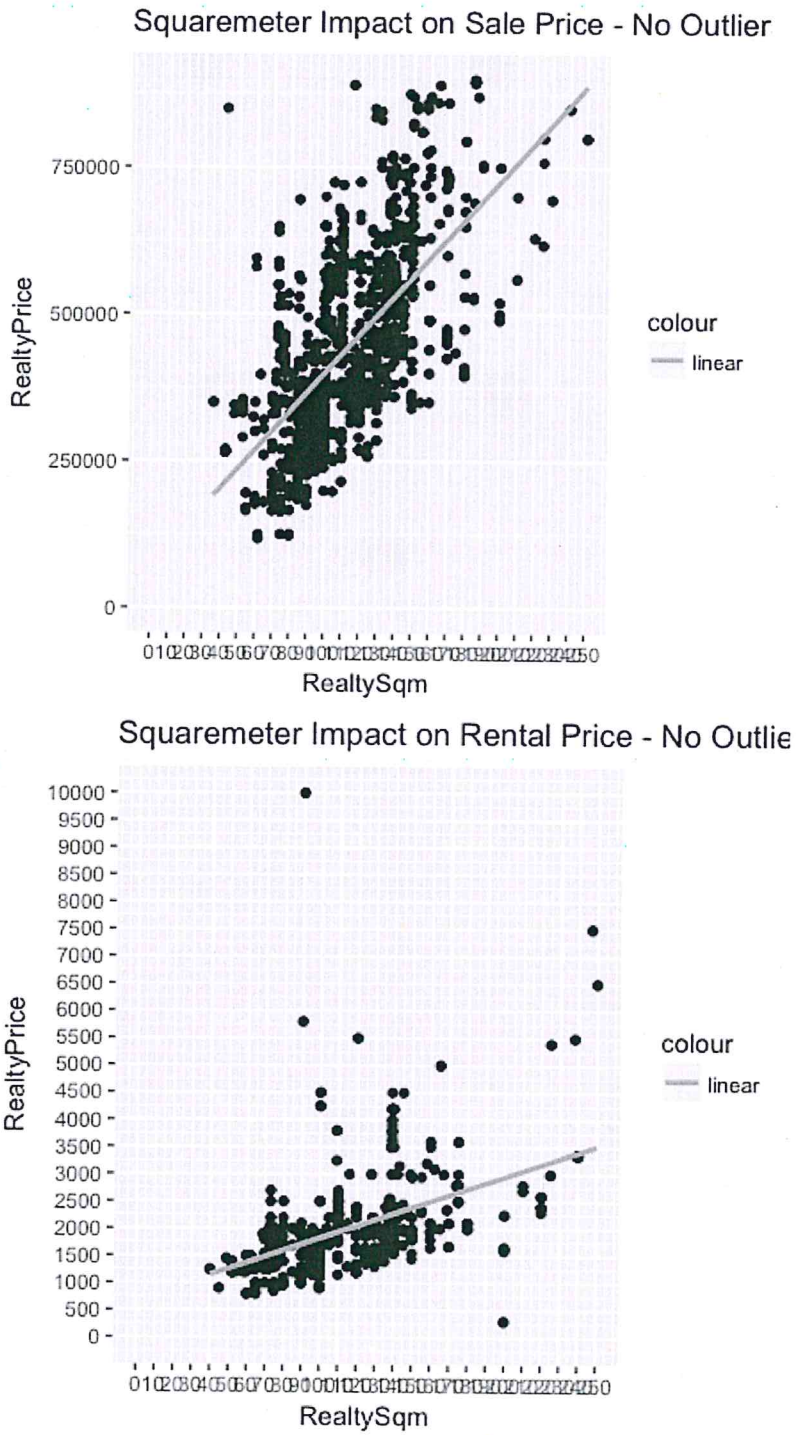
The features that are selected for model building, apart from the Realty ID and Firm ID information, are as follows:

- Price (either for sale or rental), NA/Total Ratio: 0/3137
- Area (square meter), NA/Total Ratio: 0/3137
- Age (building age), NA/Total Ratio: 227/3137
- Room Number (including bedroom and living room), NA/Total Ratio: 43/3137
- Floor (the floor building's been up to), NA/Total Ratio: 428/3137

Going back to examining adjusted R-square scores of the features, it's been observed that room number presents the highest score (0.1) followed by building age (0.04) and area (0.03). That being said, the impact varies based on the advertisements type to be either for sale or for let. For instance, the correlation score (cor.test) between area and the sale price is 0.70 while the same ratio is 0.54 for rentals. Similarly, while building age

turns out to be a significant determinant of sale price, has a much more limited impact on rentals. In Figure 2, the impact of area is given to underline these varieties (For each feature to be included or not included in the model, detailed visualizations have been done).

Figure 2: Square Meter Impact on Sale and Rental Price



It should also be noted that, in case business preferences are changed, the rule-based approach makes it quite practical to manually add new constraints or status groups. That also proves the longevity advantage of the selected approach.

2.2.2 Model Building

Following the meetings with business side and focusing on their needs and exploratory data analysis, the model is deployed. The model, which reflects rule-based approach, is consisted of five different status groups:

- **Duplicate Records:**

  Exact same value in following features: Firm ID, Price, Building Age, Area (square meter), Property Floor and Room Number

- **Near-Duplicate Records I: (changes in Floor ID)**

  Exact same value in the following features: Firm ID, Price, Building Age, Area (square meter) and Room Number

  Different value in: Floor ID

- **Near-Duplicate Records II: (+- 10% price change of houses for let)**
  - Only runs among the Rental advertisements to observe duplication.
  - Exact same value in following features: Firm ID, Building Age, Area (square meter), Property Floor, Room Number and Floor ID.
  - Different value in: Rental Price (minor transformation in the rental price of the house, the advertisements might be duplicated with reduced prices rather than updating the original entity)
- **Near-Duplicate Records III: (+- 3% price change of houses for sale)**
  - Only runs among the Sale advertisements to observe duplication.
  - Exact same value in following features: Firm ID, Building Age, Area (square meter), Property Floor, Room Number and Floor ID.
  - Different value in: Sale Price (minor transformation in the sale price of the house, the advertisements might be duplicated with reduced prices rather than updating the original entity)
- **Near-Duplicate Records IV: (+- 5 square meter change in the area of a house)**

- Exact same value in following features: Firm ID, Building Age, Area (square meter), Property Floor, Room Number and Floor ID.
- Different value in: Square meter (minor transformation in the area of the property, which may be acted as a secondary advertisement)

Each status group that is constructed signifies the possible motivations of the advertisers to resort generating duplicate or near-duplicate records. This presumptions and redefinitions are checked several times by searching the ID numbers and checking details in the original dataset. For instance, firms might be appealed to generate a brand-new entity with reduced sale price (or stating gross area rather than net one) once they saw the limited interest in the original one. The Floor ID on the other hand represents another twist. It was defined to reduce the possibility of defining records as duplicates while they were indeed different apartments in different floors but maybe in the same complex.

3. EVALUATION OF THE OUTCOMES

This section discusses the outcomes of the study by reflecting on the performance scores of the project outcome and the existing algorithm's outcome.

3.1. Evaluation of the Project Performance

The steps of project outcome and the files generated are stated below:

- First transformation – dupSusCoupleArray.csv: Imitating the pair structure of the reference document ("smiler.csv") in order to come up with bare minimum that is necessary for measured comparison. This file results all the exact duplicates and near-duplicates as pairs and returns total of 6198 rows. It is considerably higher than the first results of previous algorithm but as it will be proved before the vast difference is caused by unreliable grouping of the smiler.csv file. Before transforming it into better versions, internal distribution among the status groups are stated below:
  - Duplicates: 4196 rows
  - Near Duplicate I: 991 rows
  - Near Duplicate II: 485 rows
  - Near Duplicate III: 500 rows
  - Near Duplicate IV: 76 rows
- Second transformation - duplicatesRealtyIDs.csv: This file reduces the row number of previous outcome (pairs, a=b, a=c, b=c etc.) by bringing all the entities that are defined as duplicate/near-duplicate in the same row (a, b, c). By this way, the row number is reduced to 770 that is consisted of 2437 unique IDs. In other words, most of the duplications are entered more than two times (2437/770 = 3,16)
- Third transformation – duplicatesRealtyIDsOnlyStatus1.csv: This file makes the same transformation above by only taking the first status group (direct duplication) into consideration. The outcome returns as 670 rows and consisted of 1886 unique IDs. In other words, most of the duplications are in fact exact duplications rather than modifications, i.e. near-duplications. The same

continues however: most of the exact duplications are entered more than two times (1886/670=2,81).

- Fourth Transformation – duplicatesFirmIDs.csv: This file is based on the previous version of duplicatesRealtyIDs.csv. It is still consisted of 770 rows which all stand for 5 status groups but rather than stating the 2437 unique property IDs, they are changed with Firm IDs. Consequently, there are only 194 unique Firm IDs exists in this file ("ID1, ID2, ID3, ID4" to "FirmID1, FirmID2, FirmID1, FirmID1").
- Fifth Transformation – duplicatesFirmIDsOnlyStatus1.csv: This file does the same transformation for the exact duplicates only. The row number is 670 (same with the duplicatesRealtyIDsOnlyStatus1.csv file)
- Sixth Transformation – duplicatesFirmIDsSimplified.csv: This file represents the real outcome. The row number is obviously the same since the conditions are not changed but everything is brought together under single row: "Row Number (stating location of the Firm ID in the original dataset), Firm ID, duplicate number (repetition)".
- Seventh transformation – duplicatesFirmIDsOnlyStatus1Simplified.csv: Same simplification is done for only the exact duplications.
- Eight transformation – duplicateFirmIDsPerformance.csv: This file states the scores that is derived for each firm based on total number of entities / duplicate number of entities. It basically returns this: "Firm ID, Portfolio, Duplicates, Score".
- Ninth transformation – duplicateFirmIDsOnlyStatus1Performance.csv: This file makes the final transformation only for the exact duplicates.

3.2. Evaluation of the Existing Algorithm Performance

First of all, the existing algorithm, 'smiler.csv.' file, needs to be transformed since it did not contain the Firm ID information. Secondly, it wasn't known how the existing algorithm defined the duplicate condition. Thirdly, and most importantly, how the algorithm grouped the similar records was not stated. As a result, four different modified versions of the smiler.csv file has been developed to achieve measurable results.

- Original file: Consisted of 2485 rows, 2485 pairs

- First transformation – SmilerRealtyIDs.csv file: This file brings all the entities that was defined as duplicate by the existing algorithm. In other words, it just disrupts the pairing (a=b, a=c etc.) and presents all entities in the same row (a, b, c etc.) By this way, 2485 rows were reduced to 1201 rows. More importantly, it's been observed that smiler.csv file brought back 3137 unique ID's. In other words, the existing algorithm tries to match all the records as duplicates to each other which signifies its greatest deficiency.
- Second transformation – SmilerFirmIDs.csv: This file only builds on the previous one. Rather than stating the unique property ID's in the same row (1201 rows), it just switches property ID's with the Firm ID's that were generated by them. In a nutshell, it transformed from "a, b, c, d etc." to "FirmID1, FirmID2, FirmID3, FirmID1". Developing this file stands as an in between step.
- Third transformation - SimplifiedSmilerFirmIDs.csv: This version is the main step which paves the way for inclusion of Firm IDs. The 1201 rows of duplicates are reduced to 627 rows once the Firm ID condition is applied. It returns "Row Number, Firm ID, Duplicate Count" (627 rows is less than project outcome which is 790 rows).
- Fourth transformation – SmilerFirmIDsPerformance.csv: Similar to project's outcome, this file returns "Firm ID, Total Number of Records (Portfolio), Number of Duplicate Records, Score" for each firm in a single row. The file is consisted of 194 unique ID's and signifies as the main comparison point with the project objectives.

3.3. Conclusion

This study found out that the performance of existing algorithm is inadequate and most importantly unreliable. It wasn't able to sort out direct duplications which was stated as Status 1 group. In other words, even when further machine learning classifiers are not utilized (as further steps) to define boundaries of the sub groups or even when the status groups are not defined at all, project's direct duplication outcome is superior to the existing

one. In order to able to match the conditions of smile.csv file, exact duplications analysis are done in each and every step but still there were a significant gap.

This study returns 2437 unique IDs to be defined as duplicate or near-duplicate including all five status groups. In other words, out of 3137 entities, 2437 entities are alerted as some sort of duplication. Out of these 2437 entities, 1186 one of them are direct duplicates. The existing algorithm on the other hand, returns 3137 unique IDs among 3137 unique IDs which would be even more problematic in larger datasets.

Figure 3: Performance Score Difference Between the Project Outcome and the Existing Version

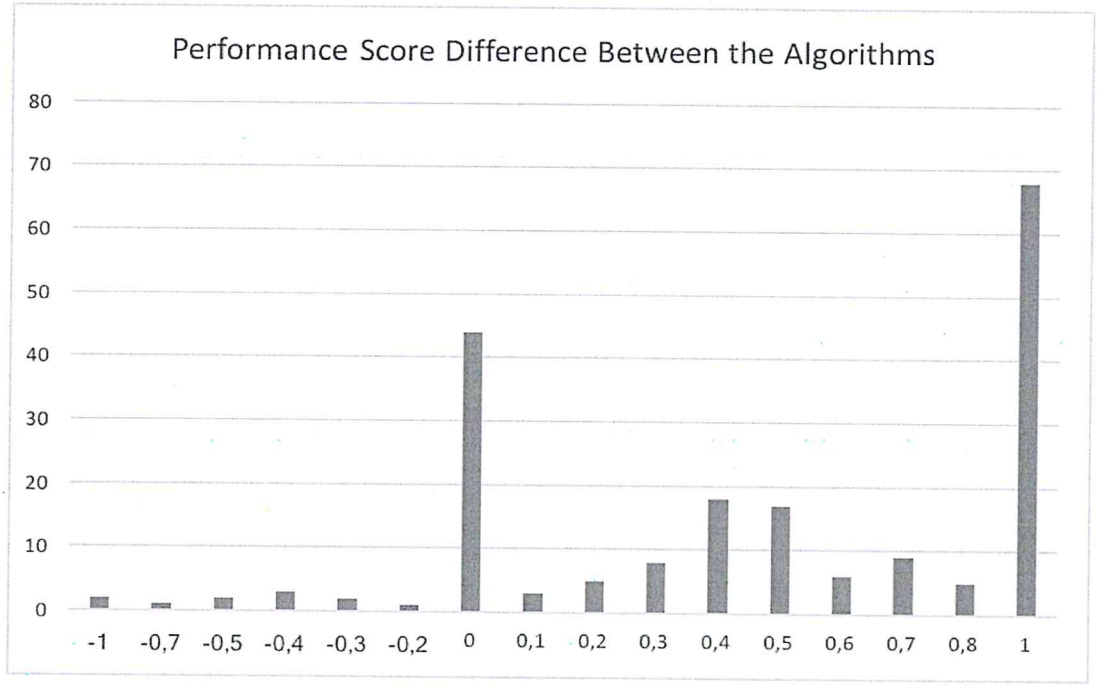
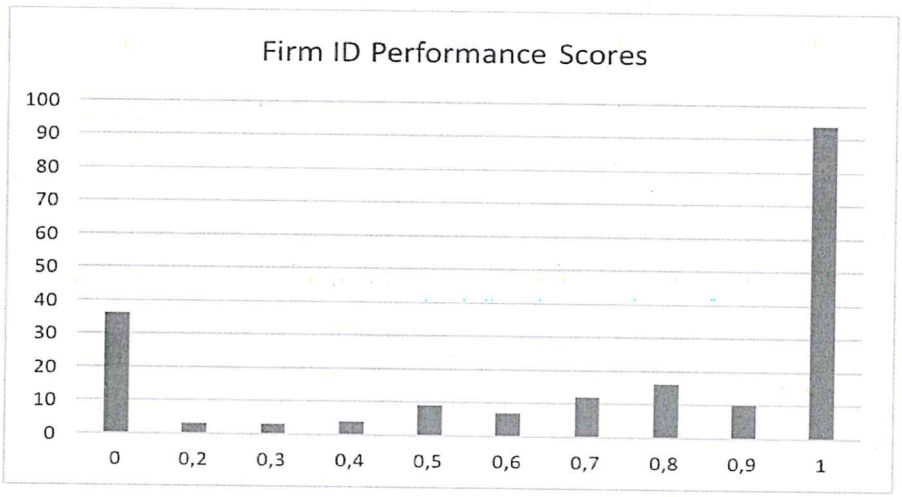


Figure 4: Firm ID Duplicate Scores (Total to Duplicate Ratio)


Firm ID Performance Scores
100
90
80
70
60
50
40
30
20
10
0
0
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

4. DELIVERED VALUE AND FURTHER STEPS

4.1. Project's Delivered Value

In this study, the duplicate record detection problem is solved through a rule-based approach to achieve utmost precision and simplicity. Duplicate and near-duplicate records are analyzed within 194 distinct sub-groups, i.e. Firm ID's, in line with the business priorities. Following the exploratory data analysis, small number of features are selected and five distinct status groups are developed. These distinct status groups provided the opportunity to review, update or modify the definition of duplication in case business priorities are changed. In the meantime, the analysis and measurement of existing algorithm's performance is presented. Taking existing algorithm's inadequacies into consideration, this study positively improves the performance of the duplicate record detection. The outcome of this project is reliable, accurate and most importantly long lasting. The formation of the five different status groups are open for update any time based on business side preferences.

4.2. Social and Ethical Aspects

The proliferation of duplicate and near-duplicate records wasn't directly classified as 'spamming' during the scope of this study. After all, entry of similar records could be unintentional as well since an alert system hasn't been activated before (an identification software/algorithm exists but it has been acted upon) and some Firm ID's could be used by multiple individuals. However, it has been observed that the ratio of total records to duplicate records was significantly higher in some cases. That reveals the possibility of some advertisers (with significantly lower ratio of duplicate records) to be compromised. Consequently, adopting the rule-based approach of this study would increase the transparency and facilitate better enforced community standards.

4.3. Further Steps

As mentioned in the previous sections of problem statement and project scope, this study has the potential of further improvement through machine learning algorithms. The five status groups are determined based on business insight and exploratory data analysis but the number of these status groups might be increased or decreased by clustering

classifiers. In addition to that, the boundaries that define those status groups (such as 3% increase or decrease in the sale price or 5 square meter raise or decline in total area) could be determined by sophisticated regression classifiers. To achieve these further steps, the dataset needs to be increased both in volume and variety. In other words, diversification of the districts or inclusion of cities would be decisive to materialize above mentioned further steps.

5. REFERENCES

- Becker, H., Naaman, M. and Gravano, L. (2010). Learning Similarity Metrics for Event Identification in Social Media. *Proceedings of WSDM 2010,* 291-300.
- Chen, Q., Zobel, J., Zhang, X. and Verspoor, K. (2016). Supervised Learning for Detection of Duplicates in Genomic Sequence Databases. *PLoS ONE,* 11(8), 1-20.
- Kaggle Featured Prediction Competition. (2017). Quora Question Pairs: Can You Identify Question Pairs That Have the Same Intent? Retrieved from: https://www.kaggle.com/c/quora-question-pairs
- Lee, K., Caverlee, J. and Webb, S. (2010). Uncovering Social Spammers: Social Honeypots and Machine Learning. *Proceedings from SIGIR 2010,* 19-23.
- Lin, Y. S., Liao, T. Y. and Lee, S. J. (2013). Detecting Near-Duplicate Documents Using Sentence-Level Features and Supervised Learning. *Expert Systems with Applications,* 40, 1467-1476.
- Martins, B. (2011). A Supervised Machine Learning Approach for Duplicate Record Detection over Gazetteer Records. In Claramunt, C., Levashkin, S. and Bertolotto, M. (Eds.), *GeoSpatial Semantics - Proceedings from GeoS 2011*, LNCS 6631, 34-51.
- Tai, K. S., Socher, R. and Manning, C. D. (2015). Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,* 1556-1566.
- Zheng, Y., Fen, X. and Xie, X. (2010). Detecting Nearly Duplicated Records in Location Datasets. *Proceedings of ACM 2010,* 1-7.