MEF UNIVERSITY

EMPLOYEE PERFORMANCE PREDICTION

Capstone Project

Barış Sivas

İSTANBUL, 2021

MEF UNIVERSITY

EMPLOYEE PERFORMANCE PREDICTION

Capstone Project

Barış Sivas

Advisor: Prof. Dr. Özgür Özlük

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: Employee Performance Prediction
Name/Last Name of the Student: Barış Sivas
Date of Thesis Defense: 06/09/2021

I hereby state that the graduation project prepared by Barış Sivas has been completed under my supervision. I accept this work as a "Graduation Project".

06/09/2021
Prof. Dr. Özgür Özlük

I hereby state that I have examined this graduation project by Barış Sivas which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

06/09/2021

Prof. Dr. Özgür Özlük
Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Barış Sivas and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Dr. Özgür Özlük	………………………..
2. Dr. Tuna Çakar	………………………..

ACADEMIC HONESTY PLEDGE

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University’s ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Barış Sivas 06 /09/2021

Name Date Signature

EXECUTIVE SUMMARY

EMPLOYEE PERFORMANCE PREDICTION

Barış Sivas

Advisor: Prof. Dr. Özgür Özlük

SEPTEMBER 2021, 34 pages

DogGo is a company that aims to provide safe and professional dog walking and grooming services to dog owners through the mobile application. Thanks to the DogGo application, dog owners and people who is employee of company and wants to walk their dogs (to be called Walkers) can meet on the same platform on the mobile application interface. The problem was determined by company that they needed to be able to accurately predict the performance of the walkers in the upcoming dog-walker matches, thus ensuring the correct dog walker match. This study will be planned to serve to this company for calculating their current walkers' performance in an accurate way. The relevant machine-learning model will first be based on the manual scoring system made by the company for the performance of existing employees, and then the model will be developed in the light of the gains obtained from this. For the performance of the model, the employees and their characteristics are important for the first time.

Key Words: Performance scoring, Prediction Future Performance, Regression Analysis, Polynomial Features, Ridge, MICE, KNN

ÖZET

ÇALIŞAN PERFORMANS TAHMİNLEMESİ

Barış Sivas

Proje Danışmanı: Prof. Dr. Özgür Özlük

EYLÜL 2021, 34 sayfa

DogGo, köpek sahiplerine mobil uygulama üzerinden güvenli ve profesyonel köpek gezdirme ve bakım hizmetleri sunmayı hedefleyen bir şirkettir. DogGo uygulaması sayesinde, köpek sahipleri ve şirket çalışanı olan ve köpeklerini gezdirmek isteyen kişiler (Walker olarak anılacaktır) mobil uygulama arayüzünde aynı platformda buluşabilmektedir. Şirketin ifade ettiği ve bu proje kapsamında çözülmesi hedeflenen temel sorun, walkerlerin gelecek olan köpek eşleşmelerindeki performanslarının önceden doğru tahmin edilebilmesi ve bu sayede doğru köpek walker eşleşmesinin sağlanabilmesi şeklinde ifade edilmiştir. Bu ihtiyaç doğrultusunda proje, firmaya mevcut walkerların performanslarının doğru bir şekilde hesaplanmasına hizmet etmek üzere planlanacaktır. İlgili makine öğrenmesi modeli ilk olarak firma tarafından mevcut çalışanların performansları için yapılan manuel puanlama sistemini temel baz alacak daha sonra buradan elde edilen kazanımlar ışığında model geliştirilecektir. Modelin performansı için ilk başta değerlendirmeye alınan çalışanlar ve özellikleri önem teşkil etmektedir.

Anahtar Kelimeler: Performans Skorlama, Çalışan Performans Tahmini, Regresyon Analizi, Ortalama Kare Hatası , Polinomsal Özellikler, Ridge, MICE, K- En Yakın Komşu

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

In recent years, especially with the improvements of internet and digital innovations there are many successful mobile application startups occurred in the world. According to the research, there are 472 million startups in the world as of 2020. Almost 1.4 million of these companies fall into the category of startup companies working in the field of technology. Again, as mentioned by the quote taken from the same research, for a startup company to be considered a technology startup, it must be able to offer technological products or services to the market. (Isp, 2020). Turkey is one of the leading countries in this field, especially in the region it is located in. In research from Sevencan (2020) as of 2020, nearly 5 billion dollars of technology export figures have been reached by nearly 6000 technology start-up companies in total. This is because there is a huge potential related with the number of mobile application users in Turkey. As stated by Gemius data, while the number of mobile users was 40.5 million in 2016, 52.3 million in 2017 and 56.9 million in 2018, this number was 60.3 million in 2019. This proves that there is an increase in mobile usage every year compared to the previous year. (Ulukan, 2021 as cited in Webrazzi 2021). It is obvious that the Covid-19 epidemic also has a leverage effect on the consumer habits of societies, especially in Turkey, in directing more online shopping. Not only the purchase of basic needs from online markets, but also various services (home, car rental, house cleaning, paint and renovation processes, transportation, pet care, etc.) received from online platforms are increasing in use. In research from World Economic Forum and The Boston Consulting Group [World Economic Forum] (2020), after the outbreak of the Covid-19 epidemic, all states and large organizations around the world had to fight this pandemic not only in the field of health, but also in meeting social needs, not interrupting production, and communication trends. With the idea from same research, technological solutions, internet and mobile applications, which have been in a rising trend in recent years, have gained a high momentum with the effect of the pandemic and emerge as the leading force in solving problems.

1.1. About Doggo Company

The trend of raising pets in the world is increasing day by day at a great pace. In line with the research report (Pet Sitting Market Size & Share | Industry Report, 2020–2027, 2020) the market size is estimated to be around $2.6 billion in 2019. Again, as based on the same research, it is predicted that the total market growth will be around 8.7% in the period between 2020 and 2027. When the chart below, which has been prepared within the scope of the relevant report, which shows similar characteristics in Europe and America, is examined, it will be noticed that the dog walking market constitutes approximately one third of the market. In order to meet the need in this area, many startup companies are established and provide services through mobile applications. For instance, Doggo is a company that aims to provide safe and professional dog walking and grooming services to dog owners through the mobile application. Thanks to the DogGo application, dog owners and people who want to walk their dogs (to be called Walkers) can meet on the same platform on the mobile application interface.

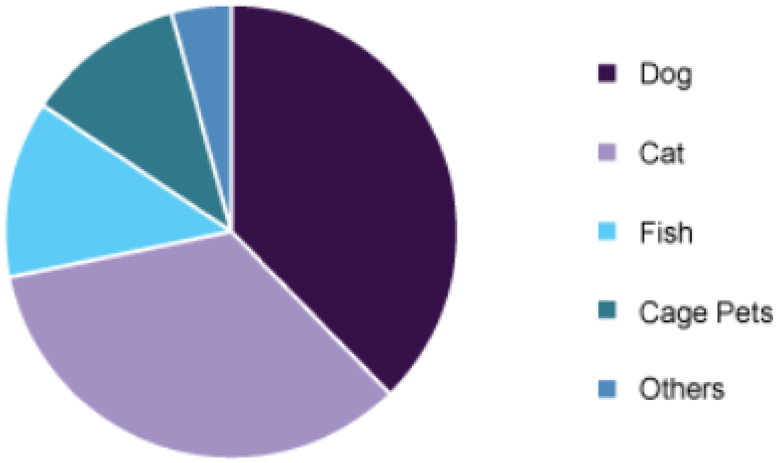


Figure 1: Global Pet Sitting Market Share, by Pet Type, 2019(%)

1.2. Employee Performance Prediction and Literature Survey

There are many cost items for companies. One of the most important one from those items is employee costs. Not only employee wages, but also operational costs have a significant place in the expense items of the company. As mentioned by the research, the increase in the rate of employee turnover almost all over the world has become another important problem for companies. Because, in addition to the above-mentioned costs, orientation trainings and start-up costs for newly arrived employees are also added. Loss of employees invested in the past can also be considered as an extra cost. (Jayadi, 2019). Since the most important reason that causes employees to leave the company is that their performance is below the company's targets, the company's selection of the right employee is an extremely important fact.

From the perspective of a different research, it was stated that the profitability of companies basically depends on how close they are to the service performance expected from them to their customers. Therefore, it was stated that the most important criterion in reaching the target of high profitability is to give the right job to the right employee (Tsekumah, 2018). According to the research from Lather et al. (2019), the success of companies cannot be separated from the performance of their employees. The researchers, conduct a study in which multiple models (Such as SVM, Random forest, naïve Bayes, Neural Networks and Logistic Regression) are used to analyze employee performance and distribute it among various classes.(p.3-6). Another research by GASTELAARS (2018) supported the idea that candidates can predict their future performance based on the scores they obtained in the interviews made during the recruitment process. He also claimed that this prediction could be made by different ML models such as linear regression, k-nearest neighbor and support vector machines.(p.2).

This capstone Project will be planned to serve to company for calculating their current walkers' performance in the accurate way. Besides of this, another aim of project is that giving to company a suitable machine learning model for helping them to predict future performance of a candidate walkers during the selection process. To make those goals real a walker-scoring model will be produced at first than it will be used to rank and score walkers that are demanded to serve for specific walking service. The expectation from this model is that to give detailed information about the walker which will lead the operation department to select the best option among demanded walkers according to their relative clusters and also to give the company a clue about a candidate walker's future performance.

1.3. About the Dataset

The data used to train the model that is planned to be created in the project consists of tables that the company has obtained and collected through the application. In the first place, an exploratory data analysis is planned to understand how the application works, how walkers interact with customers through the application, and how the service process progresses. After the preliminary preparation processes and the necessary data cleaning, the features to be used to train the model will be selected and the model will be trained by using the data for model training. Afterwards, it is expected that the next steps will be planned by evaluating the model performance on the test data together with the company. At the beginning of the project, sources shared some data via the company operation team. These datasets cover about their walker's information ("walkerparameters.csv"), their customer information ("dogparameters.csv"), and the service details "walks.csv" and "demands.csv". With a quick look at the datasets and first data cleaning processes, the first data analysis period was started. These observations will be discussed at the Exploratory Data Analysis part.

1.4. Processes Plan with the Reference of Literature Survey

As the projects and data sets get deeper, the analysis and the models to be used may differ. However, the methods considered in the first place are listed as follows: A semi-supervised learning model, which will be useful for merging differing data tables to combining those (Labelled/unlabeled data) for building a model. In research from Zhu (2012), semi-supervised learning is a kind of predictive algorithm that uses a dataset with a small amount of labeled data but mostly unlabeled data. It can be considered as a very useful salvage model, especially for problems where it would be very costly to find labeled data. To the parallel this research, at the beginning of the Doggo project, there are some examples of data labelled from domain experts via data sharing process. A polynomial regression model will be produced according to the features comes from merged tables. After getting expert knowledge for the predicted data comes from first model, the next iterations will be planned. Due to the discrete values of the "Puan" column, there could be a difference between the predicted (continuous) and actual values (discrete). In the future stages of the project, it can be considered to use the clustering method as an alternative method. As an example, in the related study (Sarker, Shamim, Zama, Rahman, 2018), data clustering and decision tree algorithms were used to estimate employee performance. The researchers first divided the performance scores into 4 different clusters with the k-means clustering algorithm, and then made the next year's performance prediction according to these groups using a decision tree (p.2.).

2. PROJECT STATEMENT AND METHODOLOGY

In this section, the definition, purpose, analysis stages (EDA) and success criteria of the project are detailed specifically. The scope of the project is explained in line with the priority criteria of the project, especially in the light of the interviews with the company that is being studied.

2.1. Problem Statement

When a dog owner creates a walking service, walkers can create walking demand from their walker app. Walker scoring model is used to rank and score walkers that are demanded to serve for specific walking service. This model will lead the operation department to select the best option among demanded walkers. Besides, in order to select best option from walker team for demanded services, first the operation team should decide which candidate should be chosen as a walker. Due to the high turnover rate, for not only best match (Walker-Demanded Service) but also high performance of the walker is crucial. According to the meeting notes with operation teams of the company, when the performance of walkers gets down, the probability of walker quitting gets higher. In this case, it causes the operational costs of the company to increase in the long run. Walker Scoring project will be also the second step of automatic matching of dog and walker.

2.1.1. Project Objectives

Within the framework of the negotiations with the company, the project output objectives can be summarized under the following 4 headings.

a) Scoring walkers eligible to serve a specific walking service,
b) Determine a base offset score for walking service
c) Dependent/Independent variable analysis
d) Performance analysis based on specific features of new comers

2.1.2. Project Scope

The scope of the project includes a general data analysis phases, which are shown as below step by, step:

a. Data Cleaning

b. Data Analysis
c. Feature Engineering
d. Data Modeling
e. Performance Evaluation
f. Data Visualization

After the meeting notes, which were done with the company, with the goal of the project, the scope of the project was determined as building a prediction machine-learning model via python libraries for candidate employees. It was decided to develop the first models by using the regression method on the performance system made for the current employees by the doggo operation teams. According to this current performance system, employees are scored between 1 and 4 (1 worst, 4 most successful). The model to be developed is planned to create an estimation algorithm based on the interaction of the employees with other information (average walking time, average feedback from the average customer, total number of walks, walking distance, etc.) kept in the system.

2.2. Methodology

Before the model and analyzes were developed in line with the project purpose and determined targets, preparatory work was carried out on the shared data set. As it was mentioned before there are five different tables (walkerparameters, dogparameters, walks, walkdemands and walkerscores) related with services in various scopes. The walkerscores table was build and shared with project team lately then other tables due to the performance operation calculations. The main objective of this project is focusing on the performance of the walkers. Therefore, in this project walkerparameters and walkerscores tables were mainly used especially for model validation part.

In the Walkerparameters table, in addition to the personal information of a walker (gender, date of entry into the system, id number, etc.), various performance indicators (average service time, customer feedback score rate, total walking time, service area) information, service type (Type A, B, C, etc.). In the Walkerscores table, besides the walker id, there is information about login to the system, the region where it is located, and the score given to the walker (on a 1-4) scale by the company's operation unit.

After merging two tables (walkerparameters & walkerscores) there are 35 rows and 820 rows occurred. With a quick look into this table, it can be said that dataset needs some preprocessing steps in order to get ready analysis.

One of the observations that attracted attention at first glance at the dataset was that there may be outliers in some columns. For this reason, it was examined how the dataset distributions of the columns were. From the distributions shown as below, it could be clearly seen that dataset consists of numerical, categorical and discrete values in various columns.

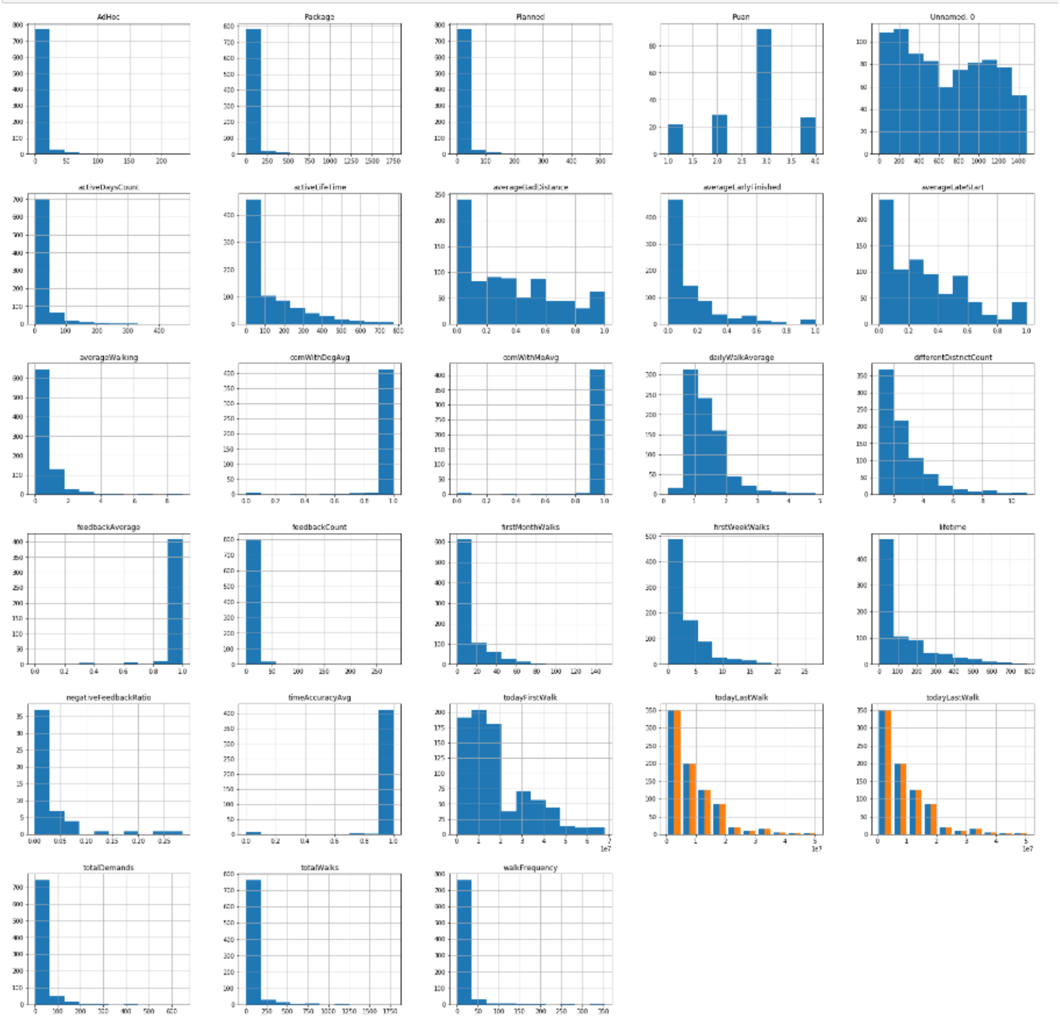


Figure 2: Distribution of the Values

The correlation analysis was done to see if there is any correlation between any of two feature or not. After the results and graph were shown, we can say that there are high correlation between some of the columns, which could be effective for analysis under the condition of all features, was used for analysis.

- Totalwalks- Active dayscount(%95)
- Activelifetime- Lifetime(%92)
- Activelifetime- todayfirstwalk(%91)

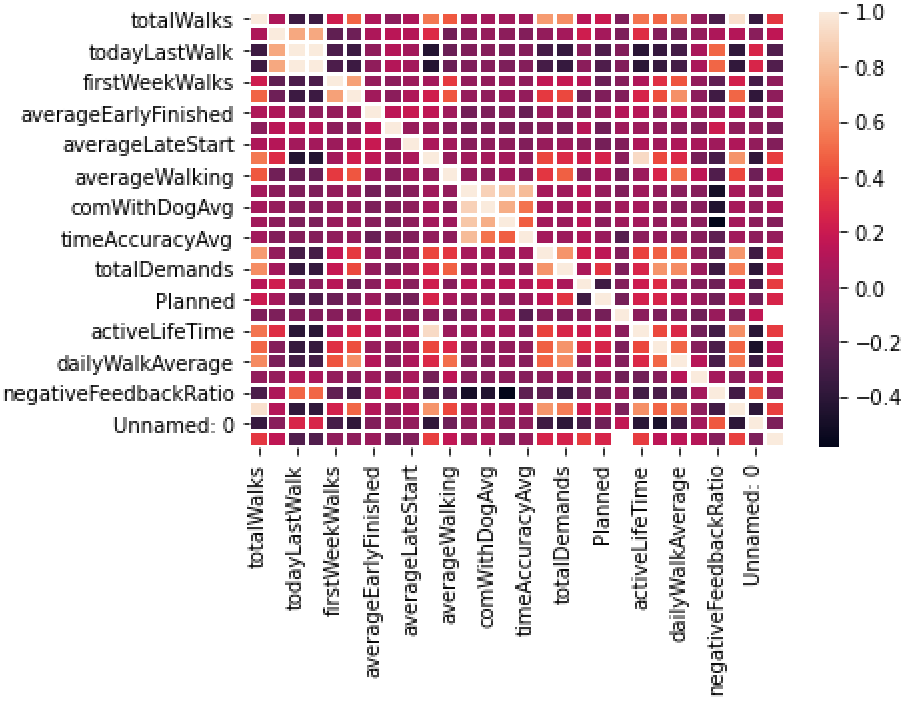


Figure 3: Correlation Matrix of Features

It was decided to use the “Puan” column from the Walkerscores table as a label in future models. Afterwards, it was decided to train the model in the light of the available data to remove the null rows in this column. In the light of the first examinations made afterwards, the following sequential preparation processes were applied to the data set, respectively.

- Some of the columns were removed because of having uniq values. (Like neighborhood, district)
- Some of the column’s (Such as gender, walkertype) datatype was generated to numeric type from the object dtype due to the regression analysis requirements.

- In the features with NaN value, mode values were given to these lines due to the discrete distributions.
- The outliers were removed to clear the features where some outliers were detected. Minimal reduction was made because the number of data was low.

2.2.1. Exploratory Data Analysis

In line with the aim and scope of the project, the main goal was to correctly identify the successful walker and to discover other features that could be parallel to its success, and the initial basic analyzes were prepared in this direction. In particular, the following sequential analyzes were conducted in order to understand the relationship between the columns in the data set and where the common features of the successful walkers are concentrated. The main messages obtained from each of them are given on the related studies. Performance estimation models, which are aimed to be created in the advanced phases of the project, will be trained with this basic feature, and then the Score values determined as the target tag will be estimated with the maximum possible accuracy and minimum error.

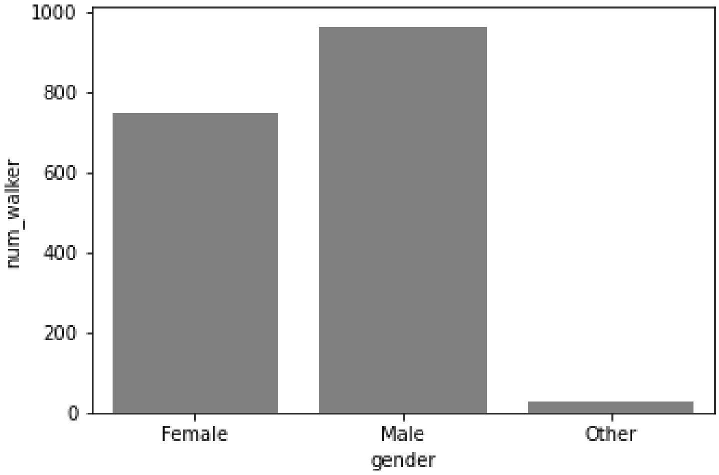


Figure 4 : Gender Distribution of Walker

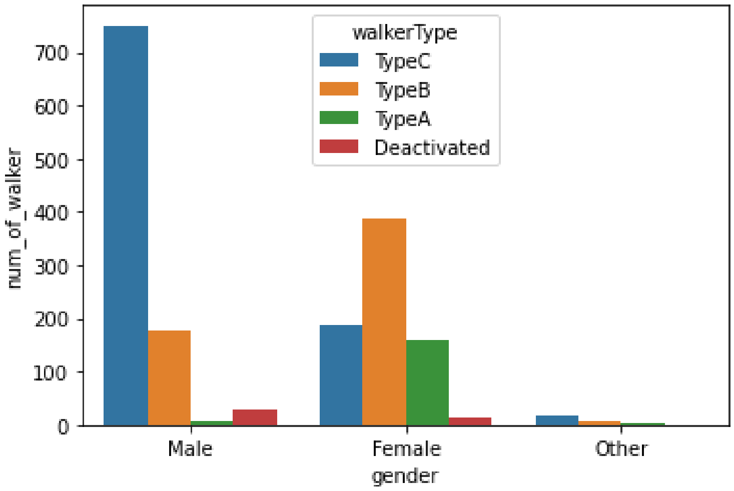
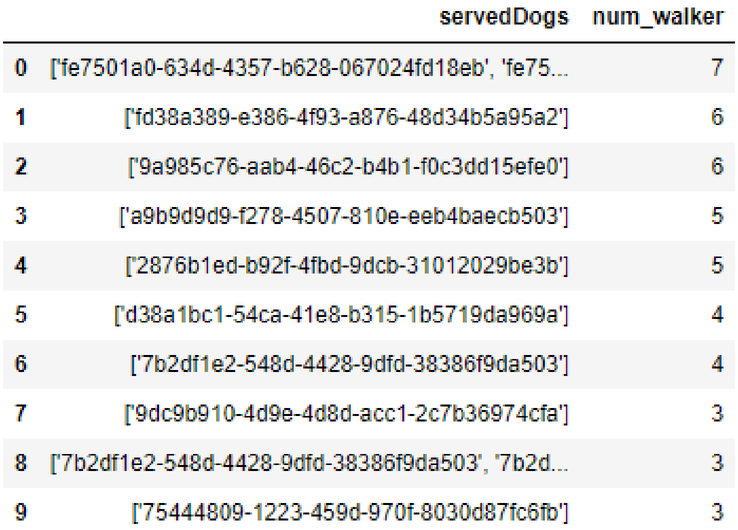


Figure 5: WalkerType Distribution with Respect to Gender

- When the distribution of the walkertypes were checked according to the gender, it was understand that the most of the walks of the males occurs by the TypeC and also the most of the TypeC walks done by male walkers. This is happen because the TypeC walks done with the dog which are over than 25 kg. (TypeA walker, can carry at most 10 kg dogs; TypeB walker can carry dogs between 10 kg- 25 kg dogs; TypeC walker, can carry dogs more than 25 kg.)
- When the dogs were ranked according to the different walkers they walked, it was seen that a dog went for a walk with 7 different walkers at most. According to the next analysis, it can be said that most customers do not want to take walking services from a different walker. It could be seen as a good impression for Doggo Company with their walker’s performances.

Table 1: Numbers of Different Walker Services Top10



	servedDogs	num_walker
0	['fe7501a0-634d-4357-b628-067024fd18eb', 'fe75...	7
1	['fd38a389-e386-4f93-a876-48d34b5a95a2']	6
2	['9a985c76-aab4-46c2-b4b1-f0c3dd15efe0']	6
3	['a9b9d9d9-f278-4507-810e-eeb4baecb503']	5
4	['2876b1ed-b92f-4fbd-9dcb-31012029be3b']	5
5	['d38a1bc1-54ca-41e8-b315-1b5719da969a']	4
6	['7b2df1e2-548d-4428-9dfd-38386f9da503']	4
7	['9dc9b910-4d9e-4d8d-acc1-2c7b36974cfa']	3
8	['7b2df1e2-548d-4428-9dfd-38386f9da503', '7b2d...	3
9	['75444809-1223-459d-970f-8030d87fc6fb']	3

- When the total number of km walked by walkers sorted according to the weather conditions, the most number of walks were occurred in cloud and rain type of weather. From the exchange of ideas with the project team on this graphic, it was thought that the dogs would be particularly badly affected by sunny weather, and therefore, the walking times and demands for walking in sunny weather would be relatively low.

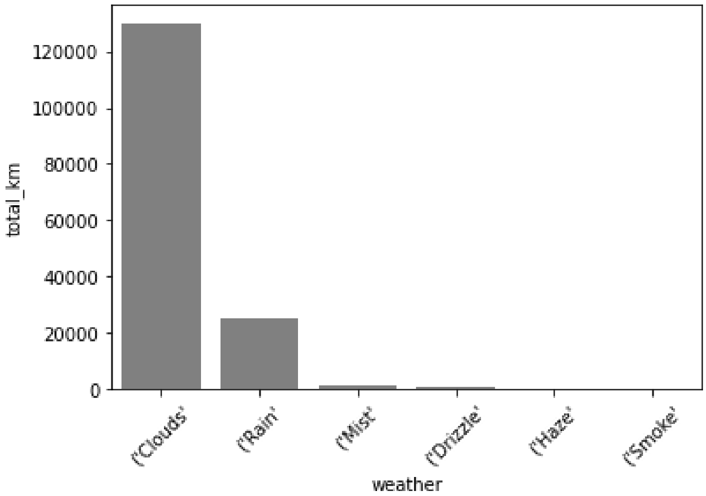


Figure 6: Number of Total Km vs. Weather Conditions

- When the walkers in the districts and the number of walks were examined, it was seen that Kadıköy. Beyoğlu and Şişli took the first three places. In the opinion of the project team in the company, the fact that it is much more common in these districts are explained by the fact that the company was first established and spread in these regions, the geopolitical location of these three regions in the city and the dense population.

- When the details of the data of the districts that take the first 5 places according to the number of walks are examined, Kadıköy, Beyoğlu and Şişli take the first three places in parallel with the total number of walks after the walkers start to work.

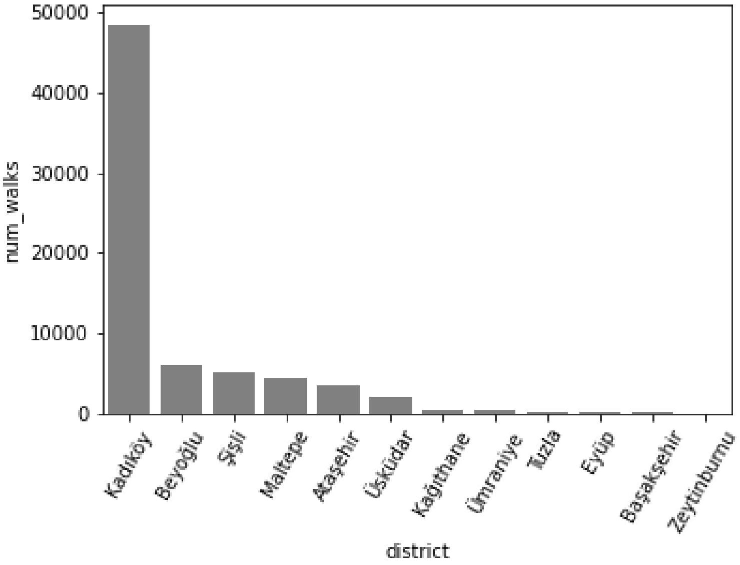
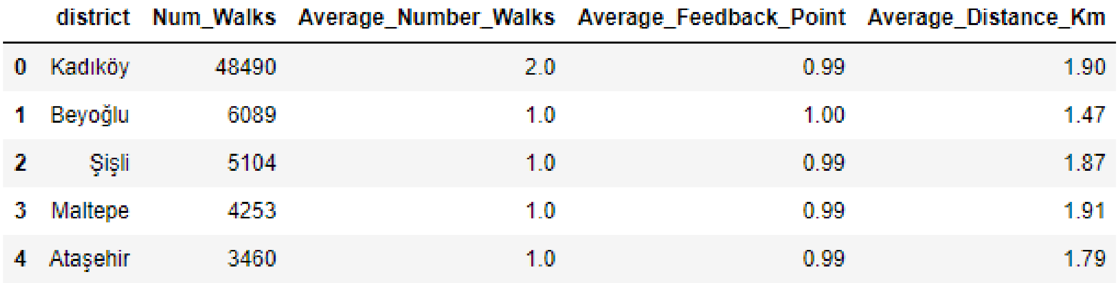


Figure 7: Distribution of Walkers by Districts

Table 2: Performance Comparison of Walks By Region

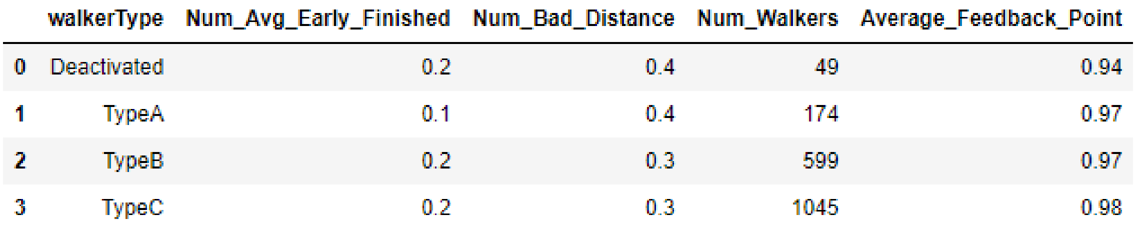


	district	Num_Walks	Average_Number_Walks	Average_Feedback_Point	Average_Distance_Km
0	Kadıköy	48490	2.0	0.99	1.90
1	Beyoğlu	6089	1.0	1.00	1.47
2	Şişli	5104	1.0	0.99	1.87
3	Maltepe	4253	1.0	0.99	1.91
4	Ataşehir	3460	1.0	0.99	1.79

- Considering the average daily walks made based on the types of walkers, it was seen that the highest rate was in type c. Then, this time, walking types were compared on the basis of success criteria such as average early completion rates, poor walking distances, average late start times, and at the end of all these results, it was seen that the most successful walker type was c. (Look at the Table 3)

- The most important result that can emerge from this is that a newcomer walker being type C can be considered as the first sign that he will most likely be a successful walker.

Table 3: Performance Comparison By Walker Type



	walkerType	Num_Avg_Early_Finished	Num_Bad_Distance	Num_Walkers	Average_Feedback_Point
0	Deactivated	0.2	0.4	49	0.94
1	TypeA	0.1	0.4	174	0.97
2	TypeB	0.2	0.3	599	0.97
3	TypeC	0.2	0.3	1045	0.98

2.2.2. Model Building

In the long-term meetings with the project team at Doggo, the most important problem was determined that the walkers achieved the least deviation from the expected performance in future customer matches. It was stated that it would be critical for the model to be created in this direction to predict the "Puan" feature, which is determined as the target label, with the least deviation. In addition, a guidance was given about which features would be more appropriate to use in the data set, which includes many features and shared by the company. These features derived from the previous services that walker gave to different customers. The project team though that these features could be beneficial for making predictions to “Puan” for future walker-customer matches’ performance results. These features and their descriptions are as follows;

a. Totalwalks: total number of walkings done by walker
b. Firstmonthwalks: number of walks done within 30 days after sign up by walker
c. Activelifetime: it is time difference between walkers' lastdemand / lastWalk and first walk
d. Differentdistrictcount: the total number of different district that walker was worked.
e. Dailywalkaverage: the mean number of walks done by walker
f. Walkfrequency: ratio of active day counts to active life time
g. Negativefeedbackratio: the ratio of negative feedback to walkers by customers
h. Activedayscount: unique number of day which walks of each walker.

Due to the fact that the Score column, which will be defined as a label in the shared data set, is less full than other features and the operation team uses integers here, an additional process is needed in the feature selection stage before the model. Since the model to be trained is a regression model, MICE and KNN methods were applied separately to fill in the missing data, considering that having continuous data would be more positive for the performance of the model.

MICE (Multiple Imputation by Chained Equations) Method:

One of the efficient way to fill missing values in a dataset is called MICE. This method is actually gained missing data from the other selected features and filled values contributions. It could be called actually different type of regression method. With the aim of taking care of missing values, this method uses a cycle which starts with dropping missing values' rows, then training a regression models with observed values. Again missing rows will be added to dataset and then these missing values will be filled via trained model's predicted values. According to new research, the MICE method of completing missing data basically creates a kind of regression model with the form of the model in accordance with the nature of the focus variable. It basically predicts missing values in the variables of a dataset, using a divide-and-conquer approach in a loop, focusing on another variable in each loop. When the focus is placed on one variable, MICE uses all other variables in the data set to estimate the deficiency in that variable. (Azur et al., 2011)

KNN (K- Nearest Neighbor) Imputation Method:

The K- Nearest Neighbor is simplest easy supervise machine learning algorithm to predict missing values in a dataset. It is commonly used to prepare dataset for classification or regression problems. In research from Monard (2013), not only discrete but also continuous attributes can be predicted by KNN method. Besides of this idea paper suggest that, KNN is easy to implement any attributes as a class which is considered as labelled features as distance metric. According to this research (Monard, 2013), when multiple missing values is a problem that should be solved, KNN method is one of the quick approach which can be treat easily. The working algorithm of KNN is based on calculating distances between nearest samples especially focused on labels and then finding related averages in order to assign to missing values. The most critical issue for using KNN is that defining the right K (number of nearest neighbors). Therefore in Doggo case, after trying several different numbers k=3 is selected at starting point.

Studies were carried out to establish the best regression model by using new data sets created separately with MICE and KNN. The created datasets were divided into 70%-30% train and test datasets and alternative regression models were studied, respectively. To give an example of these alternative models; In the first stage, Linear Regression and Ridge regularization models were established and comparisons were made on the R2 performance metric. At this point, when the performance values of the models were not found satisfactory enough, this time Polynomial regression analyzes including higher order variables were studied. At this stage, regression analyzes from the second degree to the 12th degree were performed, and then new models were tried by adding ridge and lasso regularization to these analyzes. A short excerpt from the relevant study (Bhattacharyya, 2020), for those who are not completely sure; Ridge and lasso regression techniques are two different methods used to reduce the complexity of the developed regression model and to prevent the model from memorizing the data set. As a result of the outputs obtained, separate models were selected for the KNN and MICE data sets and the results were compared in the next evaluation section.

3. EVALUATION OF THE OUTCOMES

This part is prepared for discussing performances of the two different selected models;
1- Regression model with Ridge regularization method with KNN Imputed Dataset,
2- Regression Model with Lasso regularization method with MICE Imputed Dataset

3.1. Evaluation of the Project Performance with First Model

With the new data set obtained after filling in the missing data with the KNN method, a total of twelve different models were studied, and different outputs such as the train-test performances of these models, the comparison of the estimated data with the real data, and the distance of the resulting data distribution to the targets were compared. In the last case, the parameters and various features of the model selected from these twelve different models are as follows.

- Polynomial degree of regression is defined as three
- While using Ridge regularization technique alpha value is set as 0.004.
- Train and test datasets were dived from whole dataset with ratios as %70-%30.
- The R2 train and test scores of the model was generated nearly as %40- %25 separately.
- The distribution of the predicted value versus actual value is plotted as below:

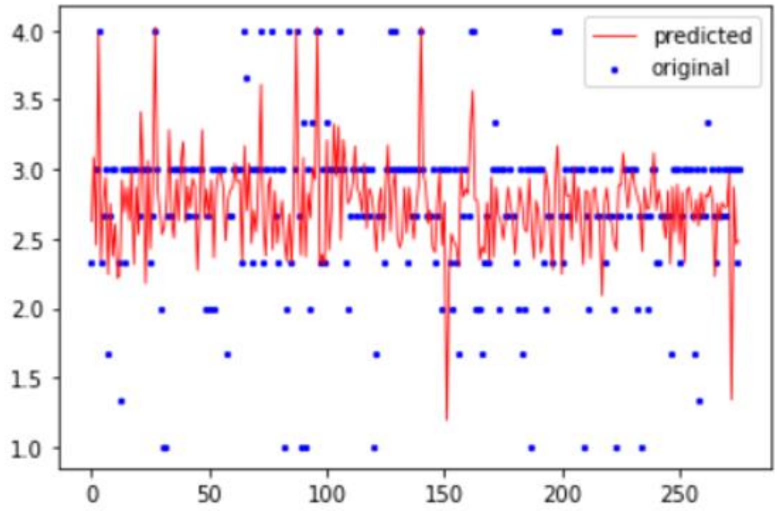


Figure 8: The Distribution of The Trained Model Outputs With KNN Dataset

3.2. Evaluation of the Project Performance with Second Model

With the new data set obtained after filling in the missing data with the MICE method, a total of twelve different models were studied, and different outputs such as the train-test performances of these models, the comparison of the estimated data with the real data, and the distance of the resulting data distribution to the targets were compared. In the last case, the parameters and various features of the model selected from these twelve different models are as follows.

- Polynomial degree of regression is defined as eleven.
- While using Lasso regularization technique alpha value is set as 0.00001.
- Train and test datasets were dived from whole dataset with ratios as %70-%30.
- The R2 train and test scores of the model was generated nearly as %63- %40 separately.
- The distribution of the predicted value versus actual value is plotted as below:

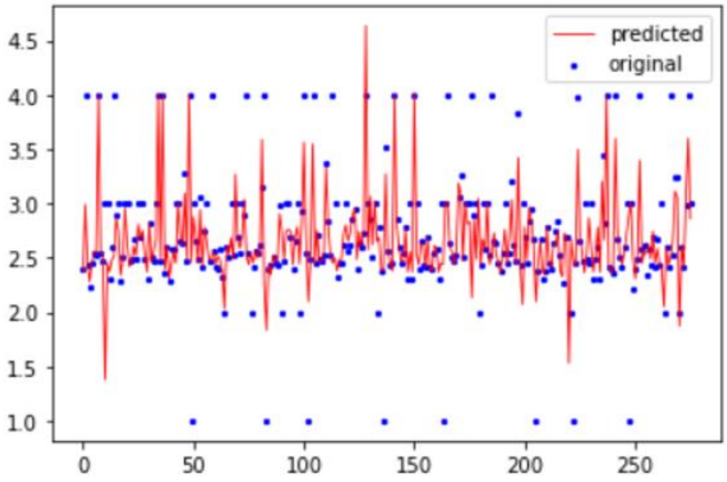


Figure 9: The Distribution of The Trained Model Outputs With MICE Dataset

3.3. Conclusion

In the last case, by using the models selected for two different data sets, it was ensured that the relevant datasets made predictions on the version of the labeled Score columns that were empty. By comparing the estimated data set with the actual values, the performances of both models were obtained in the last case. According to the obtained R2 values, respectively, 35% of the model working with KNN dataset and 56% of the model working with MICE dataset emerged. When we recalled the R2 values that emerged during the training of both data sets, KNN gave 40%-25%(train-test scores) and MICE gave 63%-40%(train-test scores) values, respectively. The final values obtained at this stage, which is a kind of validation stage, can actually be considered reasonable according to the scores obtained during the training. When we compare the two methods within themselves, both the scatterplot of the data and the R2 scores obtained can be shown to be slightly more successful than MICE KNN.

The distribution of the final model’s prediction versus actual values are plotted as below:

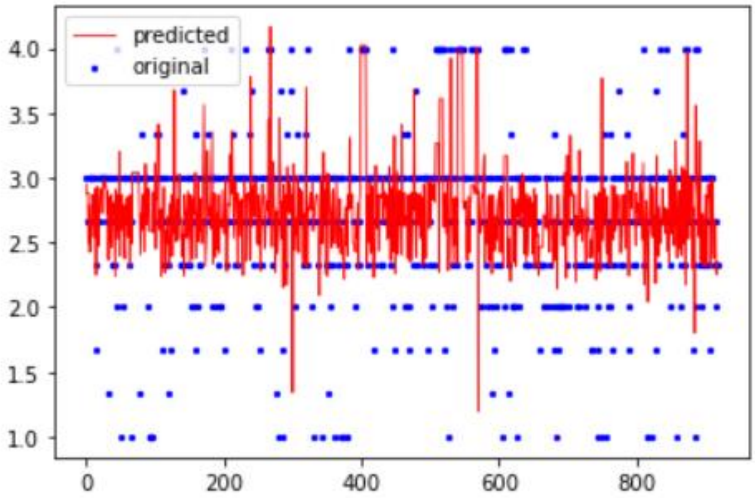


Figure 10: The Distribution of The Final Model Outputs With KNN Dataset

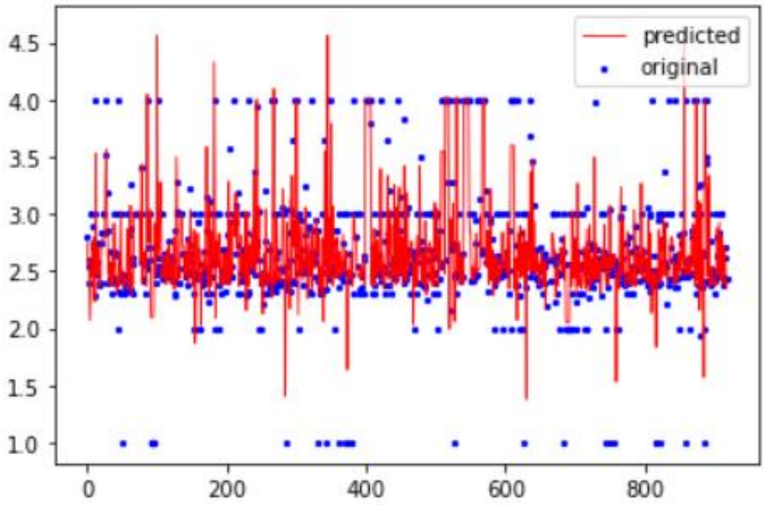


Figure 11: The Distribution of The Final Model Outputs With MICE Dataset

The distribution of the final model's prediction values are plotted as below:

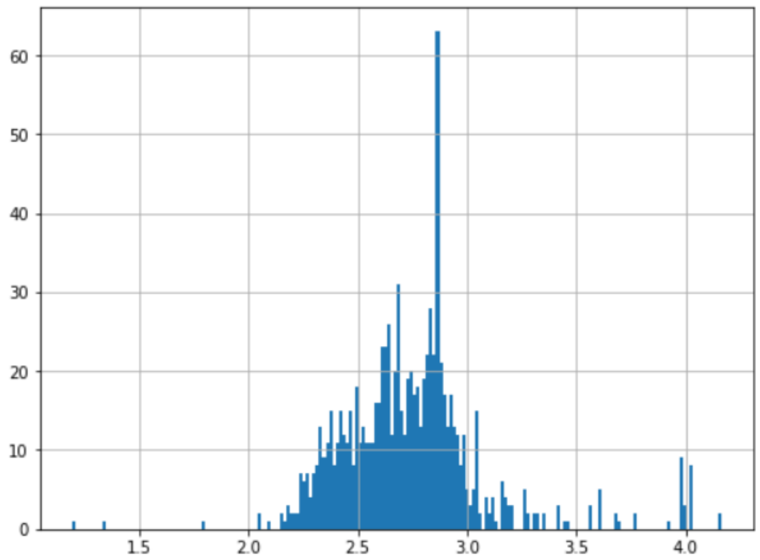


Figure 12: The Distribution of The Final Model's Prediction Values KNN Dataset

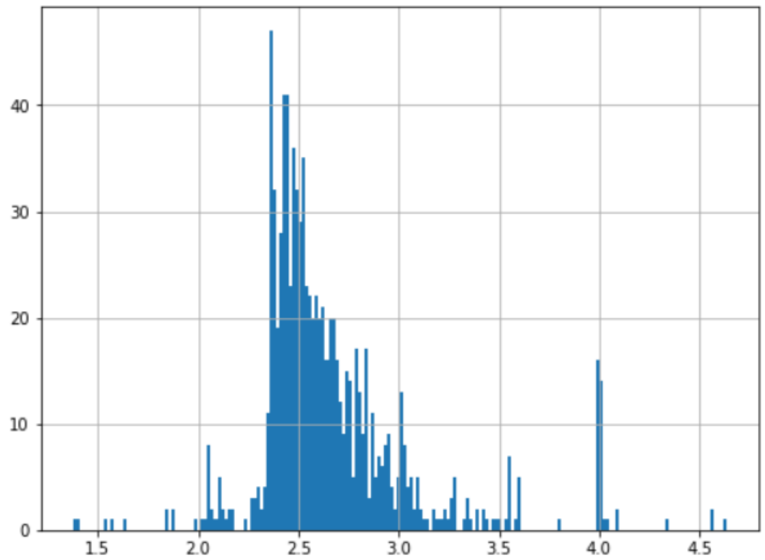


Figure 13: The Distribution of The Final Model's Prediction Values MICE Dataset

4. DELIVERED VALUE AND FURTHER STEPS

4.1. Project's Delivered Value

In this project work, it has tried to solve the problem of measuring the performance of its employees accurately so that a company can provide the most appropriate service to the relevant customer groups. Correctly solving this problem, which is very important for companies to achieve customer loyalty and employee loyalty, which are the two most important problems, is extremely critical for the future success of the company. After the study started on the datasets based on the past, tests were carried out on alternative models by selecting some of the more than twenty features and multiplying the label at hand with the help of various techniques. By using the most successful models obtained, new future cases can be solved, and I believe that new future cases will contribute to the development of the obtained model day by day.

4.2. Further Steps

The model, which I tried to summarize briefly in the previous section, will create an environment in which the model can develop further with the effect of the different scenarios that the new employees will experience in the future. In particular, the target features section, which is filled in manually at the beginning of the study, will now be able to achieve more successful results by making use of other features in the data set, with a more complex logic, with the developed machine learning model. Of course, these models developed in the future will leave their place to new techniques consisting of different parameters, and complex algorithms with more successful results will achieve performance above the current outputs. Nevertheless, I hope that this study can contribute to future studies as a starting point.

REFERENCES

Isp, N. (2020, March 25). How Many Tech Startups Are Created Each Year? - NetShop ISP. Medium. https://netshopisp.medium.com/how-many-tech-startups-are-created-each-year-27539d0a4c48

Sevencan, N. (2020, October 20). Turkey's tech startups make global waves. TRTWORLD. https://www.trtworld.com/opinion/turkey-s-tech-startups-make-global-waves-40731

Ulukan, G. (2021, March 1). Türkiye'de en çok kullanılan mobil uygulamalar ve bu uygulamaların aylık ziyaretçi sayıları. Webrazzi. https://webrazzi.com/2019/10/24/turkiye-mobil-uygulama-kullanici-sayisi-gemius/

World Economic Forum & The Boston Consulting Group [World Economic Forum]. (2020, July 1). Critical Frontier: Leveraging Technology to Combat COVID-19 [Online Forum Post].Weforum.Org. http://www3.weforum.org/docs/WEF_Critical_Frontier_Leveraging_Technology_Combat_COVID_19_2020.pdf

Pet Sitting Market Size & Share | Industry Report, 2020–2027. (2020, July). Grand View Research. https://www.grandviewresearch.com/industry-analysis/pet-sitting-market

Jayadi, R. (2019). Employee Performance Prediction using Naïve Bayes. International Journal of Advanced Trends in Computer Science and Engineering, 8(6), 3031–3035. https://doi.org/10.30534/ijatcse/2019/59862019

Tsekumah, T. (2018). How To Align Employee Targets To The Strategy. Lulu.com.

Lather, A. S., Malhotra, R., Saloni, P., Singh, P., & Mittal, S. (2019). Prediction of employee performance using machine learning techniques. Proceedings of the International Conference on Advanced Information Science and System. Published. https://doi.org/10.1145/3373477.3373696

GASTELAARS, J. (2018). Assessment scores as a predictor of your future performance and potential. Assessment Scores as a Predictor of Your Future Performance and Potential, 1. https://beta.vu.nl/nl/Images/werkstuk-gastelaars_tcm235-886119.pdf

Zhu, X. (2012, March 15). Semi-Supervised Learning Literature Survey. Minds.Wisconsin.Edu. https://minds.wisconsin.edu/handle/1793/60444

Sarker, Shamim, Zama, Rahman, A. S. M. S. M. (2018). Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm. Global Journal of Computer Science and Technology, 18(1), 1–7. https://core.ac.uk/download/pdf/231150638.pdf

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011, February 24). Multiple imputation by chained equations: what is it and how does it work? PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

Monard, M. C. (2013, June 4). A Study of K-Nearest Neighbour as an Imputation Method. ResearchGate. https://www.researchgate.net/publication/2475229_A_Study_of_K-Nearest_Neighbour_as_an_Imputation_Method

Bhattacharyya, S. (2020, September 28). Ridge and Lasso Regression: L1 and L2 Regularization. Medium. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b