DETERMINING THE COST OF A RAIL TICKET

Archana R Warrier
Department of Mathematics and Computing
Birla Institue of Technology,Mesra
Ranchi, 835217
archanarw@gmail.com

Arjita Basu
Department of Mathematics and Computing
Birla Institue of Technology,Mesra
Ranchi, 835217
arjita.basu@gmail.com

Ankit Tewari
Artificial Intelligence Engineer
Knowledge Engineering and Machine Learning Group
ankit.tewari@estudiant.upc.edu

July 7, 2019

ABSTRACT

We have used the dataset on Spanish railways which is available to the public and is available at: https://www.kaggle.com/thegurus/spanish-high-speed-rail-system-ticket-pricing. Link to the project file in GitHub is: https://github.com/archanarw/Train-pricing/blob/master/TrainTicketPricing.ipynb . Link to the project file in Kaggle is: https://www.kaggle.com/arjita2000/spanish-train-system-data-analysis/edit

1 INTRODUCTION

Spain has an extensive high-speed train network,operated by a few major operators, one of them being RENFE. Determining the price of a rail ticket of Spanish high speed railways beforehand is a challenge for travellers. It may depend on various factors such as train type, class, origin and destination city. This project aims to develop a price prediction model using linear regression and KNN to tackle above challenge.

2 DATA AND ITS PREPROCESSING

The data source for this study is the Spanish High Speed Rail tickets pricing -Renfe.The dataset includes 25,79,771 entries each with 9 features.Figure 1 shows the first five rows of the dataframe which we are going to work with.

	insert_date	origin	destination	start_date	end_date	train_type	price	train_class	fare
0	2019-04-19 05:31:43	MADRID	SEVILLA	2019-05-29 06:20:00	2019-05-29 09:16:00	AV City	38.55	Turista	Promo
1	2019-04-19 05:31:43	MADRID	SEVILLA	2019-05-29 07:00:00	2019-05-29 09:32:00	AVE	53.40	Turista	Promo
2	2019-04-19 05:31:43	MADRID	SEVILLA	2019-05-29 07:30:00	2019-05-29 09:51:00	AVE	47.30	Turista	Promo
3	2019-04-19 05:31:43	MADRID	SEVILLA	2019-05-29 08:00:00	2019-05-29 10:32:00	AVE	69.40	Preferente	Promo
4	2019-04-19 05:31:43	MADRID	SEVILLA	2019-05-29 08:30:00	2019-05-29 11:14:00	ALVIA	NaN	Turista	Promo

Figure 1: The first five rows of dataframe.

The data had 3,10,681 null values in price column, 9664 null values in train-class column and 9664 again in fare column. The null values of price column were replaced by the mean of the price column after which, the rows containing null values of fare and train-class columns were dropped. This left us with 25,70,107 rows to be evaluated.

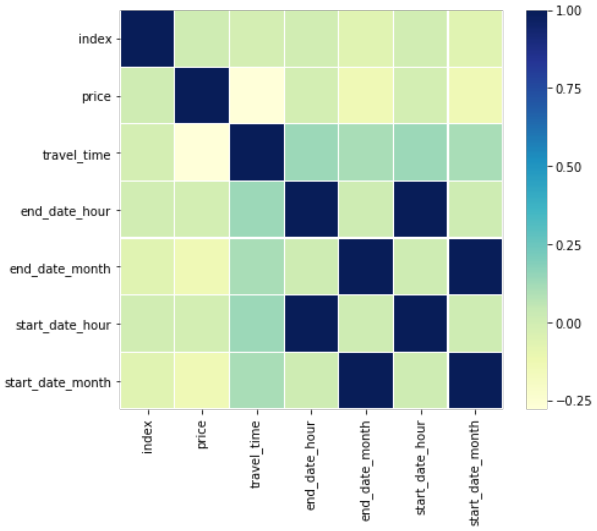

Figure 2: Correlation heatmap of the dataset

3 VISUALIZING THE DATA

Let us check some stats of the data by its visualization. There are a number of bar plots given below.

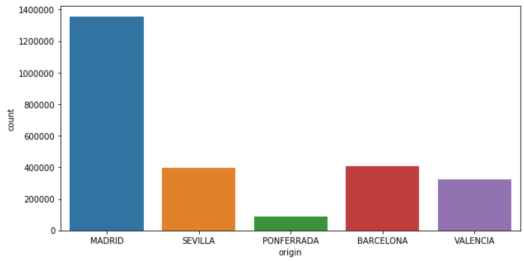

Figure 3: No. of people boarding from each station.

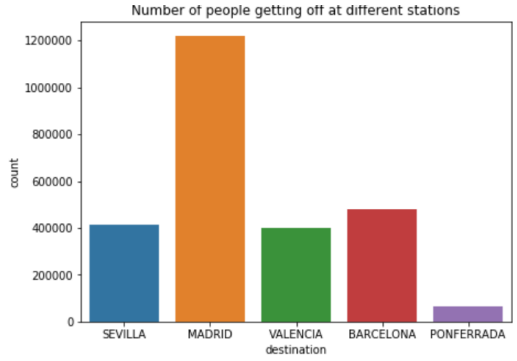

Figure 4: No. of people getting off at each station.

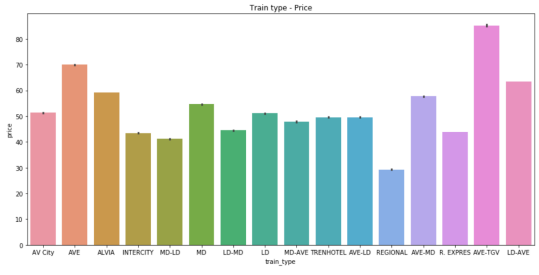

Figure 5: Train type vs. Price

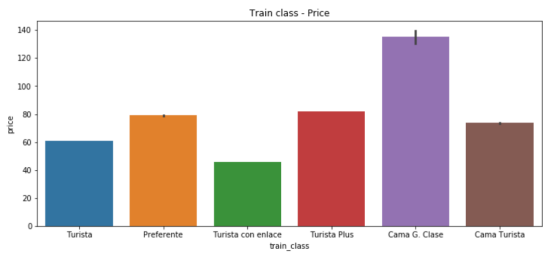

Figure 6: Train Class vs. Price

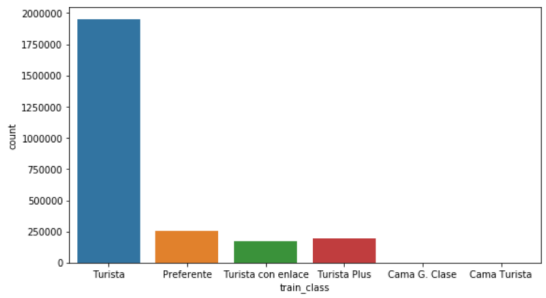

Figure 7: No. of trains of each train-class.

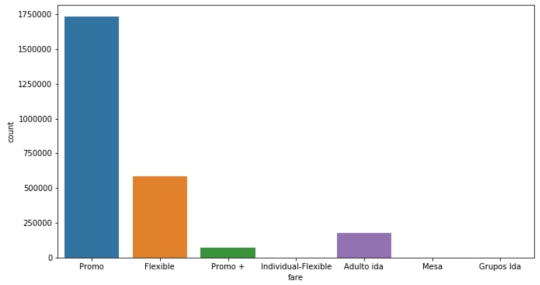

Figure 8: No. of tickets bought from each train category.

From Figure 3 we can infer that maximum number of passengers board from Madrid.Figure 4 shows us that maximum passengers get off at Madrid. Figure 5 indicates that train-type AVE-TGV is the costliest whereas Regional happens to be the cheapest. In Figure 6 we can observe that Cama G. clase is the costliest class whereas Turista con enlace is the cheapest class. Figure 7 shows us that Turista trains are the maximum in number while Figure 8 shows that promo type tickets have been sold the maximum number of times.

4 REGRESSION ANALYSIS

4.1 Linear Regression

Using the library scikit-learn, linear regression was employed on the dataset.The features considered for determining the train prices were - origin, destination, train type, fare, train class and travel time.

Most of the features were categorical, therefore to be able to apply regression, we used label encoder to transform the data into numerical format.

Here, after converting the data to numerical format and later applying linear regression, we found the linear regression score to be 0.6267. The score is found to check if the testing dataset yielded results which were similar to the expected/actual result.

4.2 KNN

KNN can be used for both regression and classification problems. The algorithm is such that the new point is assigned a value based on how closely it resembles the points in the training set. We used the algorithm with K as 5 and also found the mean squared error in predicting the prices of train tickets.

5 Conclusions

From observing the diagrams, the first conclusion we made was that the maximum number of people that boarded the trains were from Madrid and the maximum occurring destination was also Madrid.

We also see that the correlation between the start date - month and hour with the price and similarly the correlation between end date - month and hour were minimal. The journey duration had some effect on price of the tickets.

6 Acknowledgement

We would like to express our special thanks of gratitude to Ankit Tewari, who guided us on the project.Doing this study also helped us in doing a lot of Research and getting to know about so many new things. Secondly we would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited time frame.

References

[1] https://www.kaggle.com/thegurus/spanish-high-speed-rail-system-ticket-pricing

[2] https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/