TAXI FARE PREDICTION

Monalisha Ojha*
Department of Mathematics
Birla Institute of Technology
Mesra, 835215
monalishaojha974@gmail.com

Nisha Rani
Department of Mathematics
Birla Institute of Technology
Mesra, 835215
mithupriya25121970@gmail.com

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

July 10, 2019

ABSTRACT

The industry today relies heavily on data analytics to make predictions. These predictions lead to successful business models that incentivise heavily from machine learning. Popular taxi services such as Uber and Lyft provide their users with a prediction of taxi fare before the customer is mapped to a driver. We try to provide a similar solution using the open dataset provided by Kaggle.The intention is to process voluminous data in streams from Kaggle public data repository and perform different regression method and deploy a prediction engine on top of it.The key idea is to understand and implement a data analytics pipeline that forms the basis of data processing in today's software engineering.

***K*eywords** Machine Learning · Data Analytics · Linear Regression · KNN Regression

1 Introduction

Our system will process the inflow of data in order of Gigabytes from various taxi trips in the most efficient manner possible. Data processing can be further specialized here as reading the data , performing various data preprocessing tasks like data cleaning and then training a machine learning model on top of it to perform fare prediction. This project aims to develop a fare prediction model using a range of methods from linear regression to tree-based models,k nearest neighbors to tackle this challenge.

The major tasks here can be broken down into 1)Efficiently performing read operations on records in form of CSV file. 2) Processing the data in order to make it ready for consumption by the machine learning model.3) Training machine learning models and finding out the best prediction system through linear regression and knn.

Features of the locations, time of day ,longitudes and latitudes, month,week , passenger counts will be used to predict the taxi fare.

2 Related Work

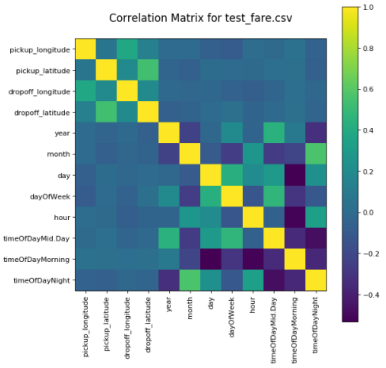
The fare of taxi ride is function of the duration of the ride (sum of drop charge, distance charge and time charge).

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

We are trying to solve a similar problem: estimating ride duration without real time data, by analysing data collected from taxis. Being able to do such estimation would help making better future predictions.

3 Dataset

The training data is consisted of 1958885 observations and 14 features. For the initial preprocessing , our team has inspected each feature of the dataset to 1) remove features with frequent and irreperable missing fields or set the missing values to zero where appropriate ,2) remove irrelevant or uninformative features or duplicate features . The team has split the data into train , validation , and test sets. Since the dataset is relatively large ,first 10 data was deemed sufficient for testing and validation sets. Consequently , several feature selection techniques were used to find the features with the most predictive values to both reduce the model variances and reduce the computation time.Based on prior housing price estimation , the first effort was manual selection of features to create a baseline for the selection process.



fare.png

Figure 1: Correlation matrix of dataset

3.1 Hypothesis Generation

The next step to solve any analytics problems is to list down a set of hypothesis, which in our case are factors that will affect the fare of taxi.

1.)Time of Travel : During peak traffic hours, the taxi fare may be higher. 2.)Day of Travel : Fare amount may differ on weekday and weekends. 3.)Pickup or Drop-off Locations: Fare may be different based on the kind of locations. 4.)Trip distance : If the distance to be traveled is more, then fare should be higher.

4 Feature Selection and Data Cleaning

Here, we will discuss various steps used to clean the data and understand the relationship between variables and use this understanding to create better features.

4.1 Distribution of fare amount

We first looked at the distribution of fare amount and found that there were 10 records where the fare was negative. Since, cost of a trip cannot be negative we removed such instances from the data.

4.2 Distribution of Geographical Features

The range of latitudes and longitudes are between -90 to 90 and -180 to 180 respectively. But in the training data set we observed latitudes and longitudes in range of (-3488.079513, 3344.459268) which is not possible. On further exploration, we also identified a set of 1649 records which had both train pickup and drop-off coordinates exceed the test pickup and drop-off coordinates.

4.3 Distribution of Trip Distance

Using the pickup and drop-off coordinates we calculate the trip distance in miles based on Euclidean Distance.One of our hypothesis was just the fare amount should ideally increase with trip distance. A scatter plot between trip distance and fare amount showed that though there is a linear relationship,the fare per mile (slope) was lower, and there were a lot of trips whose distance was greater than 50 miles, but fare was very low.

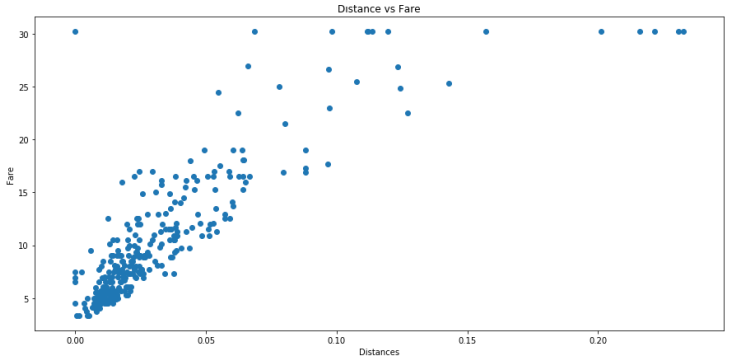


Figure 2: Distance and Fare amount

4.4 Distribution of Pickup date time

As expected, over years the average taxi fare has increased. The average fare amount at daytime is the highest while the number of trips at night are the least.

Based on the features created using this Exploratory Analysis, the baseline model using linear regression and KNN regression scored a accuracy of 83.13920724108851.

5 Methods

Linear Regression was set as a baseline model on the dataset using all of the features as model inputs. After selecting a set of features using Lasso feature selection, several machine learning models were considered in order to find the optimal one. All of the models were implemented using scikit-learn library.

5.1 Linear Regression

It is used to find a linear relationship between the target and one or more predictors. The main idea is to identify a line that best fits the data. The best fit line is the one for which the prediction error is the least. This algorithm is not very flexible, and has a very high bias. Linear Regression is also highly susceptible to outliers as it tries to minimize the sum of squared errors.The test RMSE for Linear Regression model was 4.78, and the training RMSE was 6.17.The accuracy score is 50.50139366332169

5.1.1 KNN Regression

KNN can be used for both classification and regression problems. The algorithm uses ‘feature similarity’ to predict values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. The R2 score for KNN regression model is 83.13920724108851.

5.2 Random Forest Regression

Random Forest is far more flexible than a Linear Regression model. This means lower bias, and it can fit the data better. Complex models can often memorize the underlying data and hence will not generalize well. Parameter tuning is used to avoid this problem.The Random Forest model gave an RMSE of 2.56 on test data and train RMSE of 0.014.

6 Experiments and Discussion

Mean absolute error (MAE), mean squared error (MSE) and R2 score were used to evaluate the trained models. Training (1958885 examples) and test (9914 examples) splits were used to choose the best-performing models within each category. The test set, containing 9914 examples,was used to provide an unbiased estimate of error, with the final models trained on both train and train splits. Results for the final models are provided below.

Models	MAE	RMSE	R^2 Score(in %)
Linear Regression	2.9205643127211087	4.783628153225857	50.50139366332169
KNN Regression	1.6963286975445824	2.786236253550593	83.13920724108851
Random Forest	1.5122319781477054	2.567772835575104	85.67959458833445

Figure 3: Result Table

7 Conclusions and Future work

Considering what is and what is not accounted for in the models built in this study, their predicting results are fairly accurate. To further improve the prediction accuracy, more variabilities need to be considered and modeled. This project attempts to come up with the best model for predicting the taxi fare based on a set of features including locations, longitudes and latitudes, time of the day , week , months . Machine learning techniques including Linear Regression, Tree-based models , k nearest neighbors along with feature importance analyses are employed to achieve the best results in terms of Mean Squared Error, Mean Absolute Error, and R2.The initial experimentation with the baseline model proved that the abundance of features leads to high variance and weak performance of the model on the validation set compared to the training set. This level of accuracy is a promising outcome given the heterogeneity of the dataset and the involved hidden factors , which were impossible to consider.

We have identified a couple of area where we can make improvements. Currently there are some steps that we need to perform manually. The future works on this project can include (i) studying other feature selection schemes such as Random Forest feature importance, (ii) further experimentation with neural net architectures .

8 References

[1] Machine Learning to Predict Taxi Fare — Part One : Exploratory Analysis https://github.com/atambol/taxi-fare-prediction/blob/master/Final

[2] Machine Learning to Predict Taxi Fare - Part two : Predictive Modelling https://medium.com/analytics-vidhya/machine-learning-to-predict-taxi-fare-part-two-predictive-modelling-f80461a8072e

[3] http://cs229.stanford.edu/proj2018/report/96.pdf

[4] https://scikit-learn.org/stable/, scikit-learn Machine Learning in Python.

[5] NY Taxi Fare - Comprehensive and Simple Analysis https://www.kaggle.com/danpavlov/ny-taxi-fare-comprehensive-and-simple-analysis

[6] X. Qian, S. V. Ukkusuri.: Time-of-Day Pricing in Taxi Markets. IEEE Transactions on Intelligent Transportation Systems, Vol. 18 June 2017.