***Cab Fare Prediction***

*Vikash Kumar*

*17th August 2019*

Contents

1. **Introduction**
   1. Problem Statement
   2. Data
2. **Methodology**

2.1 Pre-Processing

2.1.1 Missing value Analysis

2.1.2 Feature Selection

2.1.3 Outlier analysis

2.2 Visualization

2.3 Modeling

2.2.1 Model Selection

2.2.2 Linear Regression Model

2.2.3 Decision Tree

2.2.4 Support Vector Regression

**3. Conclusion**

3.1 Model Evaluation

3.2 Improvement

**References**

**Introduction**

***Problem Statement***

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

***Data***

* Training data:

Shape: (16067, 7)

Info:

fare\_amount 16043 non-null object

pickup\_datetime 16067 non-null object

pickup\_longitude 16067 non-null float64

pickup\_latitude 16067 non-null float64

dropoff\_longitude 16067 non-null float64

dropoff\_latitude 16067 non-null float64

passenger\_count 16012 non-null float64

* Test data:

Shape: (9914, 6)

Info:

pickup\_datetime 9914 non-null object

pickup\_longitude 9914 non-null float64

pickup\_latitude 9914 non-null float64

dropoff\_longitude 9914 non-null float64

dropoff\_latitude 9914 non-null float64

passenger\_count 9914 non-null int64

***Methodology***

***Pre-Processing***

***Missing value Analysis***

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in

certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a

proven method of resolving such issues.

In this project we would use preprocessing technique such as missing value analysis, outlier

analysis, feature selection and feature scaling.

We have seen in the data info that this data consists of missing values. We have to figure out

what is to be done with the missing data.

Methods used to impute missing data:

* Mean method : In this method we impute the missing values with the mean of the data in the column.
* Median method : In this method we impute the missing values with the median value of the column.
* KNN method : In this method we impute the missing values based on the mean of the closest neighbors.

***Feature Selection***

Feature selection is a process where we select those features which contribute most to our prediction variable or output in which we are interested in. Having irrelevant features in our data can decrease the accuracy of the models and make our model learn based on irrelevant features.

Before feature selection we have added one more variable which is ‘distance’ in km

and it calculated by the latitude and the longitude of the points in the data.

Following is the mathematical way to calculate distance between two geographical points.

R = 6373.0

dlon = radians(lon2) - radians(lon1)

dlat = radians(lat2) - radians(lat1)

a = (sin(dlat/2))\*\*2 + cos(radians(lat1)) \* cos(radians(lat2)) \* (sin(dlon/2))\*\*2

c = 2 \* atan2(sqrt(a), sqrt(1-a))

distance = R \* c

**VARIANCE INFLATION FACTOR:**

To check the multi-collinearity in our data we have calculated VIF for each variable

fare\_amount 2.925835e+00

pickup\_longitude 1.399047e+00

pickup\_latitude 1.556862e+00

dropoff\_longitude 1.441208e+00

dropoff\_latitude 1.567970e+00

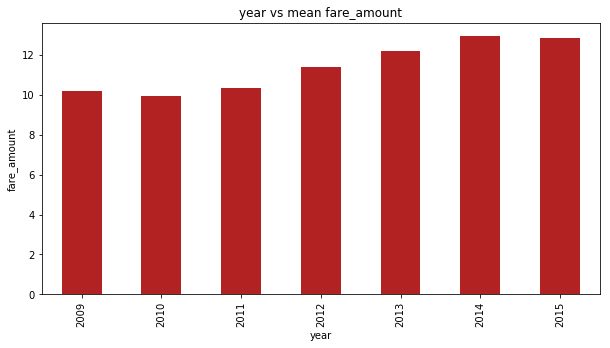
passenger\_count 1.000658e+00

distance 2.759796e+00

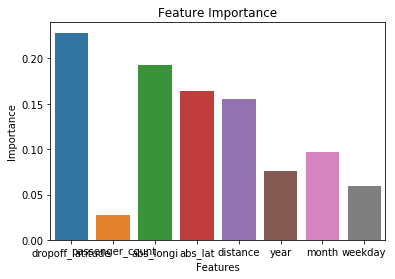
Since the VIF for every given variable is quite appropriate(<4) so we would be taking all the

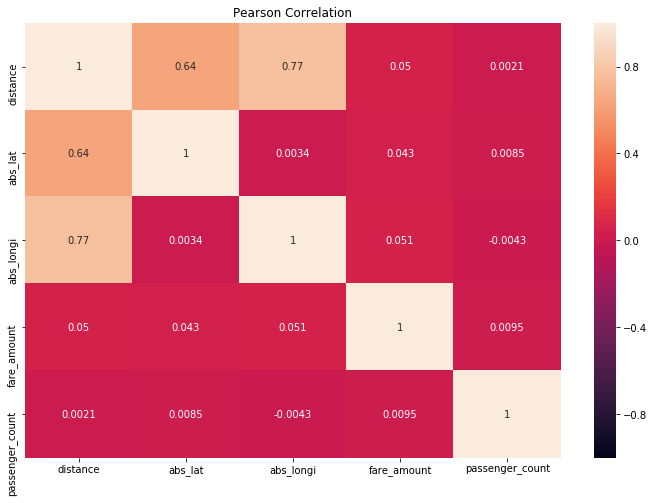
variables in our model.

***Visualization***



This shows the amount of fare in different years, by which we can make out that in which year the fare amount was high and what are the criteria which inflate the fare amount.



******

**Correlation analysis**

**Correlation analysis** is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight). This particular type of **analysis** is useful when a researcher wants to establish if there are possible connections between variables.

***Modelling***

I used different regression model to check, which model performs well and then selecting the model on the basis of their score and evaluation metrics. I have used two evaluation metrics to determine the accuracy of model i.e. R^2 square and RMSE value. To measure the performance of the regressions three standard regression metrics are used: Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Both metrics are calculated for both regressor types. For comparison RMSE is used and R^2 for parameter tuning. “The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient.”

We uses different modelling techniques to check the accuracy of the model on the given sample data. On the basis of evaluation metrics we check the model predictions and its accuracy level.

***Model evaluation***

We evaluate model on the basis of R2 score and RMSE value and then select the model as required. We got different evaluation score for each evaluating model and then we select the model which is having high accuracy and less errors in prediction.

Results obtained from Decision tree:-

Dicision tree results:

R2\_score: 0.817938

Rmse: 4.168312

MAPE results:

Out[44]:

24.5420814196481

Results obtained from Linear Regression:

Training Score : 0.6472025390345464

Validation Score : 0.7865179102507799

Cross Validation Score : 0.16434432814410044

R2\_Score : 0.5990073558559537

MAPE output: 25.527851

Results obtained from Support Vector Regression:

R^2\_Score SVR: 0.838356

RMSE SVR: 3.927638

MAPE output: 16.293515

***Conclusion***

As expected the tuning of the parameters of the regressors improves the performance. Parameter tuning with grid search can improve the performance even further after we apply different regression model on it.

Splitting the dataset and predicting casual and registered customers separately may increase the R^2 score also slightly, which is not done on this project.

A coefficient of determination of more than 80% is a decent result for the SVR regressor. All the other models performance is not satisfactory and cannot be selected in the predictions of rental bike count. We have to stick to the tuned SVR model for that as if now.

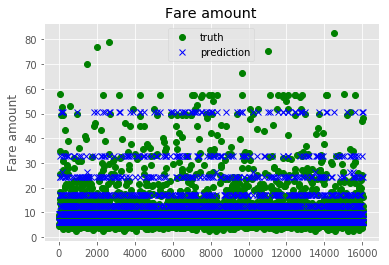
In coming future we can apply some more regression algorithm and try to calculate the accuracy and its coefficient.

***Reflection***

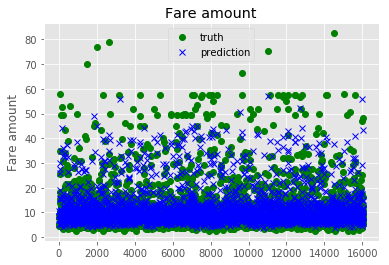
We have used different regressor model to check which model is performing well and which is having coefficient value is higher. Higher the coefficient value better the model. We have created decision tree model, linear regression model of two types and a support vector regression. We can also implement other models such as DNN regressor which is also a good model for high amount of data, which can also be used here to check if it is working fine when tuned properly.

***Visualization***

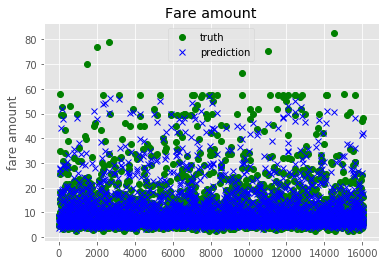
***Output of decision tree***

******

***Output of linear regression***

******

***Output of Support Vector Regression***

******

***Improvement***

The coefficient of determination of the regressors could be increased by additional iterations in training and the number of folds in the cross validation, at the expense of computing time. Of course, there are also other regressors available that might perform better on this particular dataset. For example, a wide and deep learning algorithm might be a better-performing alternative. We can also different hyperparameter tuning techniques such as randomized search and many other techniques to tune a model.

***References***

> http://www.capitalbikeshare.com

> http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

> http://freemeteo.de/wetter/

> http://dchr.dc.gov/page/holiday-schedules

> http://scikit-learn.org/stable/

> https://www.tensorflow.org

> https://www.tensorflow.org/versions/r0.9/tutorials/linear/overview.html#large-scalelinearmodels-with-tensorflow

>https://www.tensorflow.org/versions/r0.9/api\_docs/python/contrib.learn.html#DNNRegressor

> http://scikit-learn.org/stable/tutorial/machine\_learning\_map/index.html

> http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR

> http://scikit-learn.org/stable/modules/sgd.html#regression

> http://scikit-learn.org/stable/modules/svm.html

> http://scikit-learn.org/stable/auto\_examples/svm/plot\_rbf\_parameters.html

> http://scikit-learn.org/stable/modules/sgd.html#regression

> http://www.vernier.com/til/1014/

> https://github.com/tensorflow/skflow/blob/master/g3doc/api\_docs/python/

estimators.md

> http://scikit-learn.org/stable/modules/generated/

sklearn.grid\_search.GridSearchCV.html

>http://scikitlearn.org/stable/modules/generated/sklearn.grid\_search.RandomizedSearchCV.html

>http://scikitlearn.org/stable/modules/generatedsklearn.metrics.mean\_squared\_error.html#sklearn.metrics.mean\_squared\_error

> <http://scikit-learn.org/stable/modules>