Predicting Cab Fare

Himanshu Bisht

Contents

Chapter 1

Introduction

1.1 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 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.

1.2 Data

Our task is to build regression models which will predict the fare amount for a cab ride in the city. Given below is the sample of the dataset that we are using to predict the cab fare.

Number of attributes:

· pickup_datetime - timestamp value indicating when the cab ride started.
· pickup_longitude - float for longitude coordinate of where the cab ride started.
· pickup_latitude - float for latitude coordinate of where the cabs ride started.
· dropoff_longitude - float for longitude coordinate of where the cab ride ended.
· dropoff_latitude - float for latitude coordinate of where the cab ride ended.
· passenger_count - an integer indicating the number of passengers in the cab ride.

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
4.5	2009-06-15 17:26:21 UTC	-73.8443	40.72132	-73.8416	40.71228	1
16.9	2010-01-05 16:52:16 UTC	-74.016	40.7113	-73.9793	40.782	1
5.7	2011-08-18 00:35:00 UTC	-73.9827	40.76127	-73.9912	40.75056	2
7.7	2012-04-21 04:30:42 UTC	-73.9871	40.73314	-73.9916	40.75809	1
5.3	2010-03-09 07:51:00 UTC	-73.9681	40.76801	-73.9567	40.78376	1
12.1	2011-01-06 09:50:45 UTC	-74.001	40.73163	-73.9729	40.75823	1
7.5	2012-11-20 20:35:00 UTC	-73.98	40.75166	-73.9738	40.76484	1
16.5	2012-01-04 17:22:00 UTC	-73.9513	40.77414	-73.9901	40.75105	1
	2012-12-03 13:10:00 UTC	-74.0065	40.72671	-73.9931	40.73163	1

Chapter 2

Methodology

2.1.0 Pre Processing

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 Real world data are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names.

2.1.1 Missing Value Analysis

The concept of missing values is important to understand in order to successfully manage data. If the missing values are not handled properly by the researcher, then he/she may end up drawing an inaccurate inference about the data. Due to improper handling, the result obtained by the researcher will differ from ones where the missing values are present.

There are three methods which are mostly adopted to treat the missing values

1. Mean method
2. Median method
3. KNN method

Out of these three methods that method will be adopted for imputing values which will provide nearest value to the missing value.

NOTE – In case if the percentage of missing values in a particular variable is greater than 30% then it is better to remove that variable. So in my project the percentage error in variable "passenger_count" is 34.23% .So I am removing this variable from my dataset.

2.1.2 Outlier Analysis

An *Outlier* is a rare chance of occurrence within a given data set. In Data Science, an *Outlier* is an observation point that is distant from other observations. An *Outlier* may be due to variability in the measurement or it may indicate experimental error.

2.1.3 Feature Selection

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

2.1.4 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Feature scaling can vary your results a lot while using certain algorithms and have a minimal or no effect in others.

NOTE – 1.In my project I have used Normalisation which will convert the values of variable from 0 to 1.
2. Standardisation will only be use in case of uniformly distribution.

2.2 Modeling

2.2.0 Model Selection

Model selection is the process of choosing between different machine learning approaches - e.g. SVM, logistic regression, etc. - or choosing between different hyper parameters or sets of features for the same machine learning approach - e.g. deciding between the polynomial degrees/complexities for linear regression.

According to my problem statement i.e to predict the cab fare, as we can see it is clearly falls under the regression type. So for this I have used mainly two algorithm below:

1. Linear Regression
2. Decision Tree

2.2.1 Linear Regression

It is basically a statistical model. In this model instead of patterns
We are dealing with weight, coefficient or numbers of each independent variable. It is only used for regression.

2.2.2 Decision Tree

A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. A decision tree or a classification tree is a tree in which each internal (nonleaf) node is labeled with an input feature.

3 Conclusion

3.1 Model Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. There are different types of error matrices which are used to evaluate the model. So we can use two error matrices in this project:

1. MAE(Mean Absolute Error)
2. MAPE(Mean Absolute Percentage Error)
3. RMSE(Root Mean Squared Error)
   Used for time series analysis

3.1.1 Mean Absolute Percentage Error (MAPE)

For this project I have used MAPE because it will provide error in the form of percentage which can be easy to understand and to evaluate the model performance.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where A_t is the actual value and F_t is the predicted value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying by 100% makes it a percentage error.

3.1.2 Final Model Selection

Depending upon error and the accuracy we have to select a model for future data prediction.

4 Model Deployment

The concept of model deployment in data science refers to the application of a model for prediction using new data. Building a model is not generally the end the project. The model needs to be organized and presented in such a way that a customer can use it. Different methods to deploy a model in production system are following:

1. ONLINE METHOD

a. Data mining tools
-Revo Deploy R
-Orange
These tools will provide a framework or UI through which we can deploy our model at the cloud level. Based on these tools it will give UI which a non technical client can use and generate output out of it.

b. Programming languages such as JAVA, C, VB
Whenever we choose any environment such as R or Python then the library should be capable to sink with JAVA or C.

c. Database and SQL script such as TSQL, PL-SQL
Sinking the database with R or Python is also important.

d. PMML (Predictive Model Markup Language)

2. OFFLINE METHOD

Using Schedulers
In this the model should run every week and generate the output and send across to the mails.