Assignment 3

Analysis of stock pricing

|  |  |  |  |
| --- | --- | --- | --- |
| SUBJECT NUMBER & NAME | 94693 Big Data Engineering / Autumn 2021 | | |
| ASSESSMENT ITEM NUMBER/TITLE | Assignment 3 | | |
| DUE DATE | June 13 2021 11:59pm | | |
| AUTHOR DETAILS | Donovan Tay | 12964300 | Donovan.W.Tay@student.uts.edu.au |
|  |  |  |
|  |  |  |
|  |  |  |
| DELIVERABLES | Notebook:  Queries: | | |

Contents

[Business Objectives 3](#_Toc73793380)

[Business Understanding 3](#_Toc73793381)

[Data Understanding 3](#_Toc73793382)

[Data Preparation 3](#_Toc73793383)

[Data cleansing 3](#_Toc73793384)

[Model Selection and Evaluation 4](#_Toc73793385)

[Model training 4](#_Toc73793386)

[Feature importance 4](#_Toc73793387)

[Model evaluation 4](#_Toc73793388)

[Further work and areas of improvement 5](#_Toc73793389)

[Appendix 1 6](#_Toc73793390)

[Initial Data Exploration of NYC TLC 2015/16 data 6](#_Toc73793391)

# Business Objectives

The business objective of this report is to summarise the key information insights from a consolidated dataset from the New York City Taxi & Limousine (NYC TLC) and to run a predictive model on the dataset to try and predict fare amounts for trips in the New York area.

# Business Understanding

Our client would like to perform exploratory data discovery and analysis on the NYC TLC data to better understand the New York taxi market. Our team has been tasked with analysing the historical data from 2013 to 2020 and this has been

# Data Understanding

The data dictionary for the NYC TLC dataset was taken from the NYC TLC website (<https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf>). Additionally, further columns were added to facilitate the data discovery. The data dictionary is defined as:

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Expected range** |
| VendorID | A code indicating the TPEP provider that provided the record. | 1 - 2 |
| lpep\_pickup\_datetime | The date and time when the meter was engaged. Converted to timestamp. | 2015-01-01 0:00  to 2016-12-31 23:59 |

# Data Preparation

## Data cleansing

The first pass summary of the NYC Taxi 2015-6 data set is contained in Appendix 1. While the average (mean) results for the dataset seemed reasonable, upon looking at the min and max values of the dataset features, the following anomalies were noticed:

1. Drop off times outside of the expected range (2015-01-01 to 2016-12-31)
2. RateCodeID and Payment\_type values outside of expected range (1-6)
3. Passenger count had a negative mean value
4. Trip\_distance, Fare\_amount, Extra, MTA\_tax, Tip\_amount, Tolls\_amount, improvement\_surcharge, Total\_amount and Payment\_type contained negative values

To address these anomalies, the following filters were applied to the dataset:

|  |  |
| --- | --- |
| **Feature** | **Filter** |
| Lpep\_dropoff\_datetime | Between 2015-01-01 and 2017-01-10 (allow 10 days grace for long trips on New Year’s Eve 2016) |
| Lpep\_dropoff\_datetime | Remove any rows where lpep\_dropoff\_datetime is earlier than lpep\_pickup\_datetime i.e. where travel time < 0 |
| passenger\_count | Values > 0 |
| trip\_distance | Values > 0 |
| fare\_amount | Values > 0 |
| Tolls\_amount | Values > 0 |
| tip\_amount | Values > 0 |
| payment\_type | Values between 1 and 6 |
| RateCodeID | Values between 1 and 6 |
| improvement\_surcharge | Values > 0 |
| total\_amount | Values > 0 |
| Trip\_speed | The speed limit in New York State is 65mph (104kmh) – allow tolerance of 75mph (120kmh) and higher speeds considered unsustainable. |

# Model Selection and Evaluation

For this project, the Random Forest Regression model was selected as the predictive model. Random Forest model are ubiquitous and perform well across different datasets, which was requested by the client.

## Model training

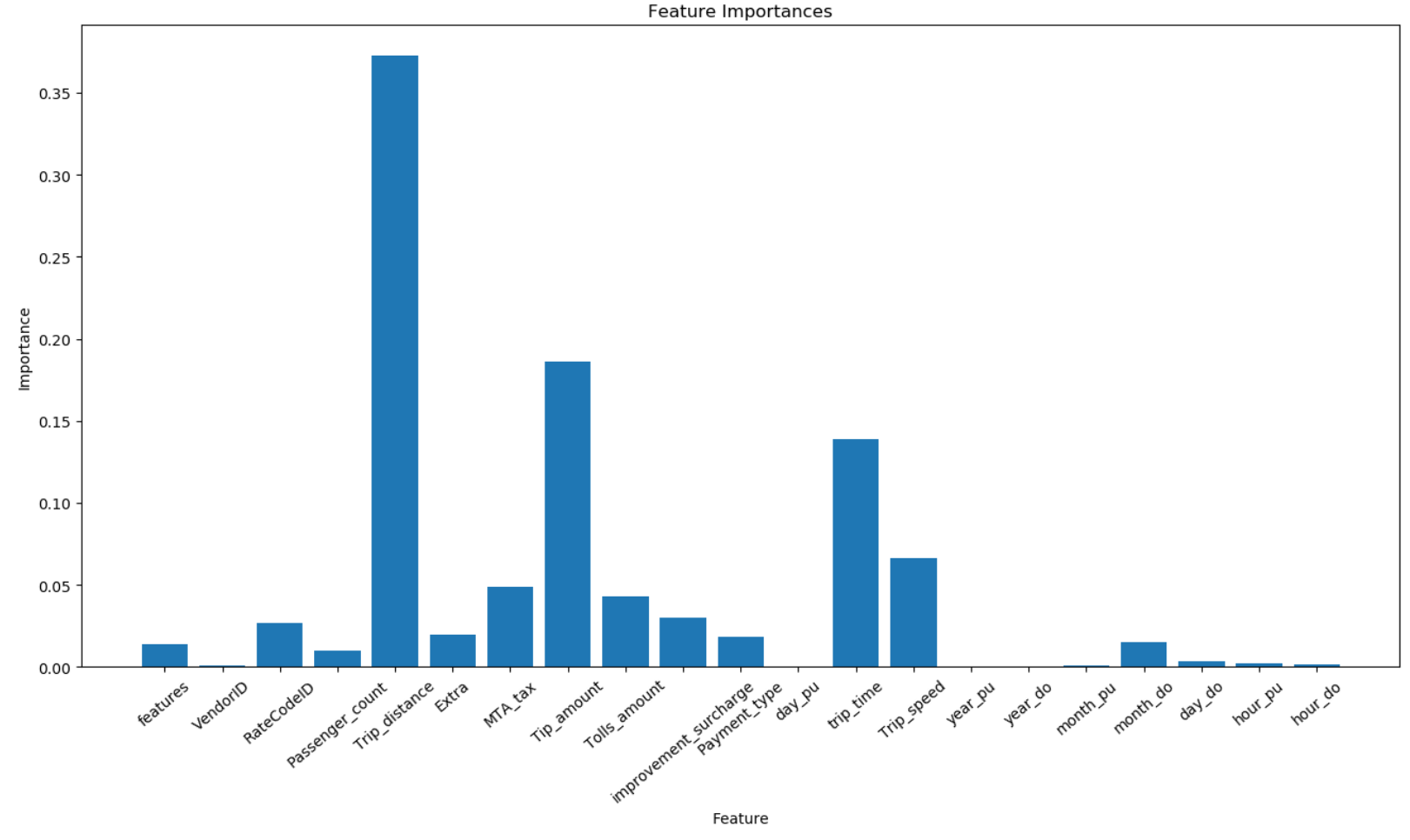
The training and test data set was divided into the following portions:

* Training – 85%
* Test – 15%

The client has requested that the model make predictions for the target class ‘total\_amount’ on the final 3 months of data, so the model has been trained on the first 21 months of data. The evaluation method selected by the client was RMSE score.

## Feature importance

From the Feature Importance plot below, the most important features for the model are passenger count, tip amount, trip time and trip speed:



## Model evaluation

Using the default Random Forest Regression model in Pyspark, the training and test results were:

* Root Mean Squared Error (RMSE) on train data = 172.98
* Root Mean Squared Error (RMSE) on test data = 5.70577

This was a reasonable result, given that there is a step-change anomaly for the last 7 months of the NYC Taxi 2015-16 dataset. The model was able to make fare predictions based on inputs, as seen in Appendix 3.

# Further work and areas of improvement

# Appendix 1

## Initial Data Exploration of NYC TLC 2015/16 data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **summary** | **VendorID** | **lpep\_pickup\_datetime** | **Lpep\_dropoff\_datetime** | **Store\_and\_fwd\_flag** | **RateCodeID** | **Pickup\_longitude** | **Pickup\_latitude** | **Dropoff\_longitude** | **Dropoff\_latitude** | **Passenger\_count** | **Trip\_distance** | **Fare\_amount** | **Extra** | **MTA\_tax** | **Tip\_amount** | **Tolls\_amount** | **improvement\_surcharge** | **Total\_amount** | **Payment\_type** |
| mean | 1.558398343 | null | null | 1.66952299 | 8.258538043 | -53.93421911 | 42.81069017 | 26.08674887 | 113.4641377 | -49.53227455 | 30.56594155 | 2.126565693 | 9.065518219 | 0.621975773 | 0.529401961 | 1.282686975 | 3.556472591 | 1.94950428 | 14.42767942 |
| min | 1 | 2015-01-01 00:00:00 | 1900-01-01 00:00:00 | 0 | -10378602.4 | -0.050000001 | -13.5518837 | -101.4903641 | 0 | -0.00824 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 | -0.1 | -0.01 | -0.31 |
| max | 3 | 2016-12-31 23:59:59 | 2253-08-23 07:56:38 | Y | 997.1 | 99 | Y | 99 | Y | 99.97 | 999.99 | 999.99 | 999.99 | 999.99 | 999.99 | 998.14 | 999.99 | 999.99 | 999.84 |