**Uber and Lyft Trip Price Prediction**

**ARTIFICIAL INTELLIGENCE**

**CS-617-A**

**UBER**

A picture containing logo

Description automatically generated

Sacred Heart University

School of Computer Science & Engineering

The Jack Welch College of Business & Technology

Submitted To:

Dr. Reza Sadeghi

Fall 2022

**Final Project Report of Uber and Lyft Trip Price Prediction**

**INTRODUCTION OF TEAM MEMBERS**

**Name:** Sanjay Kata

**SHU Email:** [katas@mail.sacredheart.edu](mailto:katas@mail.sacredheart.edu)

**Short Bio:**

I am Sanjay Kata from India. I worked as Software Developer for two years and I have hands on experience on Angular and .NET. I have good knowledge on C, C++, C#, JAVA and Python. My mini project in my undergrad is related to carpooling app & major project is related to user based collective rating system which is during my undergrad are related to Machine Learning, which is based on classification, clustering techniques for feature engineering.

**Name:** Shiva Kumar Karampuri

**SHU Email:** [karampuris@mail.sacredheart.edu](mailto:karampuris@mail.sacredheart.edu)

**Short Bio:**

I am Shiva Kumar Karampuri from India. I worked as Software Developer for four years and I have hands on experience on Java. I have good knowledge on C, C++, JAVA and Python. My projects in my work included design of HMI application which is an underground mining company.

**Name:** Payaswini Acharya Kadloor

**SHU Email:** [kadloorp@mail.sacredheart.edu](mailto:kadloorp@mail.sacredheart.edu)

**Short Bio:**

I am Payaswini Kadloor from India. I worked as Software Developer for one year and I have hands on experience on.NET. I have good knowledge on C, C++, C#, JAVA and Python. I have worked on design of BUCKLE application.

**Name:** Lavanya Ganta

**SHU Email:** [gantal@mail.sacredheart.edu](mailto:gantal@mail.sacredheart.edu)

**Short Bio:**

I am Lavanya Ganta from India. I worked as Software Developer for two years and I have hands on experience on .NET core. I have good knowledge on C, C++, C#, JAVA and Python.

**Name:** Ranadeep Adapelli

**SHU Email:** [Adapellir@mail.sacredheart.edu](mailto:Adapellir@mail.sacredheart.edu)

**Short Bio:**

I am Ranadeep Adapelli from India. I worked as .net developer for 1 year and I have hands

on experience on Angular and .NET. I have good knowledge on C, C++, C#, JAVA and Python. I have worked on a project called vehicle detection using video surveillance in my undergraduate school.

**Name:** Parupalli Bharath

**SHU Email:** [parupallib@mail.sacredheart.edu](mailto:parupallib@mail.sacredheart.edu)

**Short Bio:**

I am Parupalli Bharath from India. I worked as Full Stack Developer & Linux Admin for 3 years and I have hands on experience on Java & Unix Developer & Admin. I have been machine learning & AI Technologies in different Projects. I made electric based project using the AI interferences.

## Table of Contents

Introduction 5

* General Description
* Research Question
* GitHub Repository

Dataset Description………………………………………………………………………………………6

Related Works………………………………………………………………………………………………8

Data Exploration ………………………………………………………………………………………….9

* Univariant Analysis
* Bi Variant Analysis

Data Modeling………………………………………………………………………………….13

* Preprocessing
* Data Splitting
* Optimization
* Model Evaluation

Conclusion and Future Work…………………………………………………………….17

References………………………………….……………………………………………………17

LIST OF FIGURES

Snippet of cab\_rides.csv…………………………………………………………………………….6

Snippet of weather.csv………………………………………………………………………………6

Histogram……………………………………………………………………………………………9

Distplot……………………………………………………………………………………………..11

Pairplot…………………………...………………………………………………………………...12

Countplot…………………………………………………………………………………………...12

Linear Regression…………………………………………………………………………………..14

Random Forest Classifier…………………………………………………………………………..15

Decision Tree Classifier……………………………………………………………………………16

Lasso Regression…………………………………………………………………………………...16

Comparison of all regressions……………………………………………………………………...16

INTRODUCTION

* + **General Description:**

Sometimes it is just convenient to leave the driving to someone else. This is just less stress; more mind space and one get to use that time to do other things. Well, that is one of the ideas that grew and later became the concept behind Uber and Lyft.

Unlike public transit, the cost of an Uber or Lyft ride varies. The demand for and supply of transportation at any one time have a big impact on them. So, what exactly drives this demand? The first thing that comes to mind is the time of day; as most people commute to work or home between 9 am and 5 pm, these hours ought to see the biggest increases. As more people should take rides when it's raining or snowing, the weather could also be an influence.

Uber and Lyft's ride prices are not constant like public transport. After our data exploration and building the models, we would like to examine how distance, time, and weather impact the price estimation of a ride. Also, help the customers to choose between Uber and Lyft for their trip.

* + **Research Question:**

Which factors play a role in predicting the prices of Uber and Lyft?

* + **Github Repository:**

The code and the documentation that we have worked on can be found in the following link:

<https://github.com/parupallib/uber-lyft-prediction>

## 

**DATA DESCRIPTION**

* + **URL of the dataset:**

The data set can be downloaded from the following:

<https://www.kaggle.com/datasets/ravi72munde/uber-lyft-cab-prices>

* + **When where and how the data is collected:**

The Data set is sourced from Kaggle. There are two datasets:

**cab\_rides.csv**: This dataset gives information like distance traveled, cab type, source and destination, price of the ride, surge multiplier, timestamp, and name of the cab about trips in Boston for one week.

The cab rides dataset has more than 650K records. The data is approx. for a week of Nov '18. This interval was chosen to query the maximum amount of data possible without unnecessary redundancy. The Cab ride data covers various types of cabs for Uber & Lyft and their price for the given location. You can also find if there was any surge in the price during that time.

Below is a snippet of the data:

Graphical user interface, text, application

Description automatically generated

**Fig 1: Snippet of cab\_rides.csv**

**weather.csv:** Contains weather attributes like temperature, humidity, pressure, rain, and timestamp. Below is the Snippet of the data:

Graphical user interface

Description automatically generated with low confidence

**Fig 2: Snippet of weather.csv**

* + **The name, definition, and characteristics of features**

|  |  |
| --- | --- |
| **Weather.csv** | **Cab\_rides.csv** |
| Temp: Temperature in F | Distance: The Distance Between Source and Destination |
| Location: Location Name | Cab\_type: Uber or Lyft |
| Clouds: Clouds | Time\_stamp: Epoch Time When Data was Queried |
| Pressure: Pressure in MB | Destination: Destination of the Ride |
| Rain: Rain in Inches for the Last Hr. | Source: The Starting Point of the Ride |
| Time\_stamp: Epoch Time When Row Data was Collected | Price: Price Estimate for the Ride in USD |
| Humidity: Humidity in % | Surge\_multiplier: The Multiplier by Which Price was Increased, Default 1 |
| Wind: Wind Speed in MPH | Id: Unique Identifier |
|  | Product\_id: Uber/Lyft Identifier for Cab-type |
|  | Name: Visible Type of the Cab Eg: Uber Pool, Uberxl |

Below is the column information for the two datasets:

**Source of the code:**

**https://github.com/alfafimel/PREDICTION-OF-CAB-PRICES-UBER-LYFT-**

**RELATED WORKS**

* Ride-sourcing companies such as Uber and Lyft represent a popular and growing mode of transit in cites worldwide. These companies employ surge pricing in real time to balance the needs of both drivers and riders. The prediction of surge prices in the next few minutes to hours encapsulates the complex evolution of service fees and service demand in the short term. Surge pricing, if effectively predicted and disseminated to both drivers and riders, can be used to more efficiently allocate vehicles, save users money and time, and provide profitable insight to drivers, which ultimately helps the efficiency and reliability of transportation networks. This paper explores the spatial-temporal correlations between the urban environment, trafc fow characteristics, and surge multipliers. We propose a general framework for predicting the short-term evolution of surge multipliers in real-time using a log-linear model with L1 regularization, coupled with pattern clustering.

<https://par.nsf.gov/servlets/purl/10137166>

* The potential of an efficient ride-sharing scheme to significantly reduce traffic congestion, lower emission level, and drivers’ stress, as well as facilitating the introduction of *smart cities* has been widely demonstrated in recent years. Furthermore, ride sharing can be implemented within a sound economic regime through the involvement of commercial services that creates a win-win for all parties (e.g., *Uber*, *Lyft* or *Sidecar*). This positive thrust however is faced with several delaying factors, one of which is the volatility and unpredictability of the potential benefit (or utilization) of ridesharing at different times, and in different places. Better understanding of ride-sharing dynamics can help policy makers and urban planners in increase the city’s “ride-sharing friendliness” by designing new ride-sharing oriented systems, as well as by providing ride-sharing service operators better tools to optimize their services.

<https://www.hindawi.com/journals/jat/2019/6125798/>

* At Uber, magical customer experiences depend on accurate arrival time predictions (ETAs). We use ETAs to calculate fares, estimate pickup times, match riders to drivers, plan deliveries, and more. Traditional routing engines compute ETAs by dividing up the road network into small road segments represented by weighted edges in a graph. They use shortest-path algorithms to find the best path through the graph and add up the weights to derive an ETA. But as we all know, the map is not the terrain: a road graph is just a model, and it can’t perfectly capture conditions on the ground.

<https://www.uber.com/blog/deepeta-how-uber-predicts-arrival-times/>

**Source of the project:**

<https://ucladatares.medium.com/uber-vs-lyft-how-do-they-decide-their-price-92222122f482>

**DATA EXPLORATION**

Univariant analysis:

“The term **univariate analysis**refers to the analysis of one variable. You can remember this because the prefix (“uni”) means (“one”).” [3]

Generally, the most effective method to use univariant analysis is by plotting a histogram and distplot.

Histogram:

“A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size. In the most common form of histogram, the [independent variable](https://www.techtarget.com/whatis/definition/independent-variable) is plotted along the horizontal axis and the [dependent variable](https://www.techtarget.com/whatis/definition/dependent-variable) is plotted along the vertical axis. The data appears as colored or shaded rectangles of variable area.”[4]

Chart

Description automatically generated

Chart

Description automatically generated

**Fig 3:** Histogram

The above figure depicts the variance of each column in the data set individually.

Distplot:

“A **Distplot**or distribution plot, depicts the variation in the data distribution.”[13]

Chart, histogram

Description automatically generated

**Fig 4:** Distplot

The figure above depicts the change in the distribution of data.

Bivariant analysis:

“Bivariate analysis lets you study the relationship that exists between two variables. This has a lot of use in real life. It helps to find out if there is an association between the variables and if yes then what is the strength of association.”[5]

pairplot and countplot can be used for bivariant analysis

Pairplot:

“Pairplot visualizes given data to find the relationship between them where the variables can be continuous or categorical.”[6]

The below figure shows a relation between quality and each column of the data set.

A picture containing table

Description automatically generated

A picture containing calendar

Description automatically generated

**Fig 5:** Pairplot

CountPlot:

“A countplot is kind of like a histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of category.”[7]

Chart, histogram

Description automatically generated

**Fig 6:** Countplot

DATA MODEL

Preprocessing:

Pre-Processing techniques includes Data Cleaning, Data Normalization, Data Reduction, Data Transformation.

* Data Cleaning: “Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.” [17]
* Data Normalization: “Database normalization or database normalisation (see [spelling differences](https://en.wikipedia.org/wiki/American_and_British_English_spelling_differences#-ise,_-ize_(-isation,_-ization))) is the process of structuring a [relational database](https://en.wikipedia.org/wiki/Relational_database) in accordance with a series of so-called [normal forms](https://en.wikipedia.org/wiki/Database_normalization#Normal_forms) in order to reduce [data redundancy](https://en.wikipedia.org/wiki/Data_redundancy) and improve [data integrity](https://en.wikipedia.org/wiki/Data_integrity). It was first proposed by [British](https://en.wikipedia.org/wiki/British_people) [computer scientist](https://en.wikipedia.org/wiki/Computer_scientist) [Edgar F. Codd](https://en.wikipedia.org/wiki/Edgar_F._Codd) as part of his [relational model](https://en.wikipedia.org/wiki/Relational_model).” [18]
* Data Reduction: “Data reduction is the process of reducing the amount of capacity required to store data. Data reduction can increase storage efficiency and reduce costs. Storage vendors will often describe storage capacity in terms of raw capacity and effective capacity, which refers to data after the reduction.” [19]
* Data Transformation: “Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system. Data transformation is a component of most [data integration](https://www.talend.com/resources/what-is-data-integration/) and [data management](https://www.talend.com/resources/what-is-data-management/) tasks, such as data wrangling and [data warehousing](https://www.talend.com/resources/what-is-data-warehouse/).” [20]

Data Splitting:

“Data splitting is commonly used in machine learning to split data into a train, test, or validation set. This approach allows us to find the model hyper-parameter andestimate the generalization performance.” [21]

Linear Regression:

“Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.” [15]

Text

Description automatically generated

**Fig 7: Linear Regression**

Random Forest Classifier:

Random forest is a *Supervised Machine Learning Algorithm* that is *used widely in Classification and Regression problems*. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.[9]

Text

Description automatically generated

**Fig 8:** Random Forest Classifier

Decision Tree Classifier:

 Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

[10].

Text

Description automatically generated

**Fig 9:** Decision Tree Classifier

Lasso Regression:

“**Lasso regression** is a type of linearregressionthat uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.” [16]

Text

Description automatically generated

**Fig 10: Lasso Regression**

Here We are comparing all the classifiers, to know the best model.

Chart, bar chart

Description automatically generated

**Fig 11:** Comparison of all the classifiers

Linear regression has highest accuracy for this model.

**MODEL EVALUATION**

“Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.” [22]

Here I have used Random Forest classifier to build a new model and evaluate the performance of the model

**Random forest classifier**:

It is an ensemble learning method that operates by constructing multitudes of decision trees at training time.

**Decision Tree classifier:**

“Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes.**” [23]

**CONCLUSION AND FUTURE WORK**

Our work here predicts the price in different conditions. In future we would like to have hypothesis testing like the change in price according to temperature and change in price according to time.

**REFERENCES**

1.<https://en.wikipedia.org/wiki/Data_pre-processing>

2. <https://www.tableau.com/learn/articles/what-is-data-cleaning>

3. <https://www.statology.org/univariate-analysis/>

4. <https://www.techtarget.com/searchsoftwarequality/definition/histogram>

5. <https://www.jigsawacademy.com/blogs/business-analytics/bivariate-analysis/>

6. <https://medium.com/analytics-vidhya/pairplot-visualization-16325cd725e6#:~:text=Pairplot%20visualizes%20given%20data%20to,attractive%20and%20informative%20statistical%20graphics>.

7. <http://www.learningaboutelectronics.com/Articles/How-to-create-a-countplot-in-seaborn-with-Python.php#:~:text=A%20countplot%20is%20kind%20of,a%20certain%20type%20of%20category>.

8. <https://scikit-learn.org/stable/modules/svm.html>

9. <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random%20forest%20is%20a%20Supervised,average%20in%20case%20of%20regression>.

10 . <https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html>

11. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761#:~:text=Summary,that%20data%20in%20use%20grows>.

12. <https://www.analyticsvidhya.com/blog/2021/03/machine-learning-with-python-gaussian-naive-bayes/>

13. <https://www.journaldev.com/39993/seaborn-distplot#:~:text=A%20Distplot%20or%20distribution%20plot,with%20different%20variations%20in%20it>.

14. <https://amarchinthevines.org/2014/09/13/wine-is-the-most-healthful-and-most-hygienic-of-beverages-pasteur/>

15. <https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable>.

16. <https://www.statisticshowto.com/lasso-regression/>

17. <https://www.tableau.com/learn/articles/what-is-data-cleaning#:~:text=tools%20and%20software-,What%20is%20data%20cleaning%3F,to%20be%20duplicated%20or%20mislabeled>.

18. <https://en.wikipedia.org/wiki/Database_normalization>

19. <https://www.techtarget.com/searchdatabackup/definition/data-reduction#:~:text=Data%20reduction%20is%20the%20process,to%20data%20after%20the%20reduction>.

20. <https://www.talend.com/resources/data-transformation-defined/#:~:text=Data%20transformation%20is%20the%20process,data%20wrangling%20and%20data%20warehousing>.

21. <https://www.diva-portal.org/smash/get/diva2:1506870/FULLTEXT01.pdf>

22. <https://www.dominodatalab.com>

23. [https://www.saedsayad.com/decision\_tree.htm#:~:text=Decision%20tree%20builds%20](https://www.saedsayad.com/decision_tree.htm" \l ":~:text=Decision%20tree%20builds%20)

[classification%20or,decision%20nodes%20and%20leaf%20nodes.](https://www.saedsayad.com/decision_tree.htm" \l ":~:text=Decision%20tree%20builds%20)