INT 254
FUNDAMENTALS OF MACHINE LEARNING

PROJECT REPORT

ON

OLA BIKE RIDE REQUEST DEMAND FORECAST

Submitted by		
NAME	REG. NO.	ROLL NO.
Karma Tashi	12204002	RKM068A03
Shalini Tigga	12205663	RKM068B28
Arpit Srivastav	12205735	RKM068B29
Prakhyat Singh	12218463	RKM068B44

Under the guidance of

DR DHANPRATAP SINGH (25706)

School of Computer Science and Engineering
Lovely Profession University, Jalandhar,
Punjab, India

TABLE OF CONTENTS

DECLARATION

I hereby declare that the project work entitled “Ola Ride Request Forecast” is an authentic record of my own work carried out as requirements of Project for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of Dr Dhanpratap Singh, during January to May 2024. All the information furnished in this project report is based on my own intensive work and is genuine.

Prakhyat Singh
12218463
4th April 2024

CERTIFICATE

This is to certify that the declaration statement made by this student is correct to the best of my knowledge and belief. He has completed this Project under my guidance and Supervision. The present work is the result of his original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfilment of the conditions for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara.

Dr Dhanpratap Singh
School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab
Date: 4th April, 2024

ACKNOWLEDGEMENT

It is with my immense gratitude that I acknowledge the support and help of my Professor, Dr Dhanpratap Singh, who has always encouraged me into this research. Without his continuous guidance and persistent help, this project would not have been a success for me. I am grateful to the Lovely Professional University, Punjab and the department of Computer Science without which this project would have not been an achievement. I also thank my family and friends, for their endless love and support throughout my life.

ABSTRACT

Ride-sharing platforms like Ola have revolutionized urban transportation by providing convenient, efficient, and cost-effective alternatives to traditional taxi services.

In this project, we propose a machine-learning approach to forecast ride requests on the Ola platform. Accurate prediction of ride demand is crucial for optimizing fleet management, resource allocation, and overall service efficiency. Leveraging historical ride data, weather information, time-series analysis, and advanced machine-learning algorithms, we aim to develop a robust forecasting model capable of predicting ride requests with high accuracy and reliability.

Our methodology involves data preprocessing, feature engineering, model selection, and evaluation to identify the most effective forecasting approach. By accurately predicting ride demand, Ola can enhance service availability, reduce waiting times for passengers, and optimize driver utilization, ultimately leading to improved customer satisfaction and operational efficiency. The results of this study offer valuable insights for ride-sharing platforms seeking to optimize their operations through data-driven forecasting techniques.

Based on our research, Random Forest Regressor performs the best by giving the least error(Validation error = 7.79, Training error = 3.00).

INTRODUCTION

- The ride-hailing (Ola) service sector has been expanding for a few years, and it is anticipated to continue expanding in near future. Ola drivers must decide where to wait for passengers since they may arrive rapidly. Additionally, passengers like an immediate bike service whenever required. People who have issues with booking Ola bikes, which sometimes cannot be fulfilled or the wait time for the arrival of the trip is particularly lengthy owing to the lack of a nearby Ola bike. If you successfully reserve an Ola bike in one go, consider yourself fortunate. Ola is acquiring a greater market share and significance in a variety of transportation markets. Big data technologies and algorithms should be employed to handle the enormous amounts of information that are available to enhance service efficiency. This will allow for more accurate estimates of efficiency as well as assistance in meeting the needs of riders. This work develops a model to forecast supply and demand mismatches using information from the leading ride-hailing company in Bangalore. The percentage of Indians who travel by taxi, bus, or rail is among the highest in the world and few of the Indians 1.4 million residents own automobiles. The leading ride-hailing business in Bangalore, Ola, handles more than 1 lakh rides daily and gathers more than 5GB of data.

- It has become important for Ola (and other e-haling) company to forecast the demand for their Ola bikes so that they may better understand that demand and maximize the efficiency of their fleet management. A novel model based on users' ride request dataset is proposed to address these problems; it would include characteristics such as ride booking time, season, and weather, temp, humidity, windspeed, number of non-registered user rentals initiated, number of registered user rentals initiated, number of ride request raised on the app for that hour. This model will try to predict ride-request for a particular hour using machine learning, assisting the business in maximizing the density of Ola bikes to meet consumer demand.

METHODOLOGY

Ride hailing companies (such as Ola) are losing money and market share to their competitors, due to their failure to satisfy the trip demands of many consumers. To solve this issue, a novel model is presented out to predict ride-request for a particular hour using machine learning.

The Used Algorithms

1. Linear Regression

Linear regression is a fundamental statistical technique used to understand the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables X and the dependent variable Y, and seeks to find the best-fitting linear equation that describes this relationship.

The basic form of a linear regression model with one independent variable is given by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y is the dependent variable,
- X is the independent variable,
- β_0 is the intercept (the value of Y when X =0),
- β_1 is the slope (the change in Y for a unit change in X),
- ϵ is the error term, representing the difference between the observed and predicted values of Y

The goal of linear regression is to estimate the values of β_0 and β_1 that minimize the *sum of squared errors* (SSE) between the observed values of Y and the values predicted by the model.

Linear regression is widely used in various fields, including economics, finance, biology, and social sciences, for tasks such as predicting sales, analyzing trends, and understanding the impact of variables on an outcome.

Advantage: Linear regression is computationally efficient, especially for large datasets, and can be implemented quickly.

Disadvantage: If too many independent variables are included in the model, it can lead to overfitting, where the model performs well on the training data but poorly on new data.

2. Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that adds a penalty term to the standard linear regression objective function. This penalty term is the sum of the absolute values of the coefficients multiplied by a regularization parameter (α), which controls the strength of the penalty. The objective function for Lasso is given by:

$$\text{minimize } (\text{RSS} + \alpha \textstyle\sum_{j=1}^{p} |\beta_j|)$$

where:

- RSS is the residual sum of squares, the sum of the squared differences between the predicted and actual values,
- p is the number of features,

- β_j are the coefficients of the linear regression model for each feature j
- α is the regularization parameter.

The main idea behind Lasso is that by adding this penalty term, the model is encouraged to select only the most important features and to shrink the coefficients of less important features to zero. This leads to a sparse model, where only a subset of the features are used in the final model.

Advantage: Lasso has the ability to perform feature selection, which can improve the interpretability of the model and reduce overfitting, especially when dealing with datasets with a large number of features

Disadvantage: Lasso tends to select only one feature from a group of highly correlated features, which can lead to instability in the selected features.

3. RandomForestRegressor

RandomForestRegressor is an ensemble learning method that belongs to the family of decision tree algorithms. It is used for regression tasks, where the goal is to predict a continuous value. RandomForestRegressor builds multiple decision trees during training and outputs the average prediction of the individual trees, which improves the accuracy and robustness of the model.

Advantages: RandomForestRegressor tends to have higher accuracy compared to single decision trees, as it reduces overfitting by averaging the predictions of multiple trees.

Disadvantages: RandomForestRegressor has several hyperparameters that need to be tuned, such as the number of trees in the forest and the maximum depth of the trees, which can be challenging.

4. Ridge

Ridge regression is a linear regression technique that adds a penalty term to the standard linear regression objective function. This penalty term is the sum of the squared values of the coefficients multiplied by a regularization parameter (α), which controls the strength of the penalty. The objective function for ridge regression is given by:

$$\text{minimize } (\mathrm{RSS} + \alpha \sum_{j=1}^{p} \beta_j{}^2)$$

where:

- RSS is the residual sum of squares, the sum of the squared differences between the predicted and actual values,
- p is the number of features,
- β_j are the coefficients of the linear regression model for each feature j
- α is the regularization parameter.

The main idea behind ridge regression is that by adding this penalty term, the model is encouraged to shrink the coefficients of less important features towards zero, but not exactly to zero. This can help prevent overfitting, especially when dealing with datasets with a large number of features or features that are highly correlated.

Advantage: Its ability to reduce the impact of multicollinearity (high correlation between features) in the dataset. By shrinking the coefficients of correlated features, ridge regression can improve the stability and interpretability of the model.

Disadvantage: It is a linear model, so it may not perform well when the relationship between the features and the target variable is highly non-linear.

Dataset

The data set used in this study was a ride request dataset. This dataset would have the following attributes: ride booking time, season, and weather, temp, humidity, windspeed, number of non-registered user rentals initiated, number of registered user rentals initiated, number of ride request raised on the app for that hour. Explanation for the column names in the dataset and their values is as follows:

Season-

1. spring
2. summer
3. fall
4. Winter

Weather-

1. Clear, Few clouds, Partly cloudy, Partly cloudy
2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

casual – number of non-registered user rentals initiated.

registered – number of registered user rentals initiated.

count – number of ride request raised on the app for that hour.

Data Preparation

There are times when multiple features are provided in the same feature or we must derive some features from the existing ones. We will also try to include some extra features in our dataset so, that we can derive some interesting insights from the data we have. Also, if the features derived are meaningful then they become a deciding factor in increasing the model's accuracy significantly.

Exploratory Data Analysis

EDA is an approach to analysing the data using visual techniques. It is used to discover trends, and patterns, or to check assumptions with the help of statistical summaries and graphical representations. We will add some features to our dataset using some assumptions. And will also check what are the relations between different features with the target feature.

Model Training

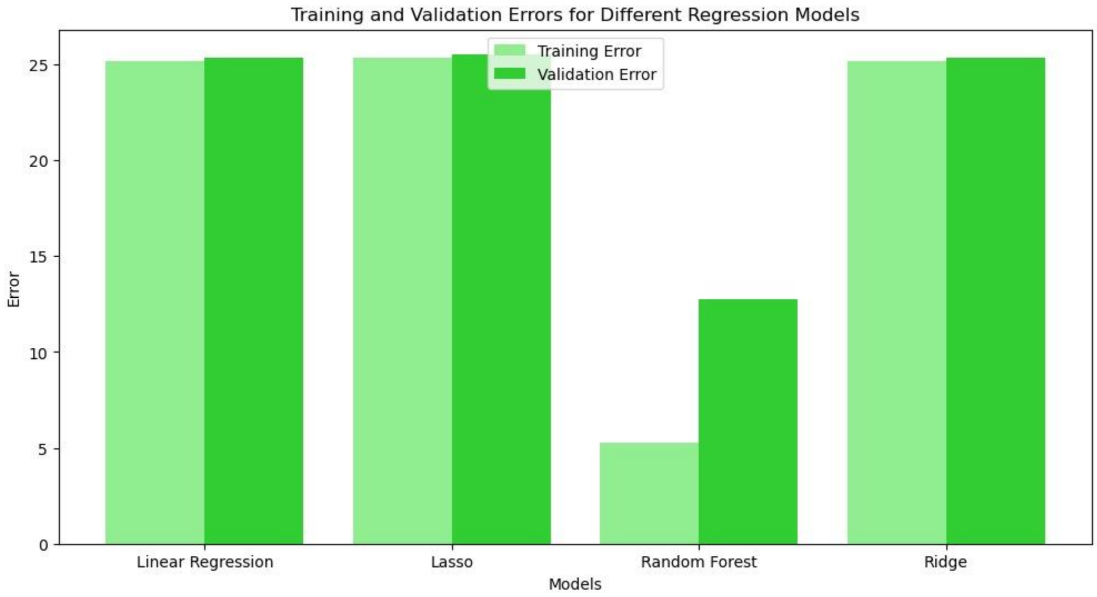
Will separate the features and target variables and split them into training and the testing data by using which I will select the model which is performing best on the validation data.

I will split the data into training and validation data also the normalization of the data will be done. I will train some state-of-the-art machine learning models and select the best out of them using the validation dataset.

DISCUSSION

Error

	Training Error	**Validation Error**
Linear Regression	20.54	20.03
Lasso	20.24	19.78
RandomForestRegressor	3.00	7.79
Ridge	20.64	20.14



As we can see Random Forest Regressor has the lowest Training and Validation Error.

CONCLUSION

In this study, we compared several machine learning algorithms and found that the RandomForestRegressor algorithm outperformed others on our dataset.

RandomForestRegressor has the lowest Training Error (3.0069790034471953) among all the models, indicating that it has the best performance on the training data. This suggests that RandomForestRegressor is able to capture the underlying patterns in the training data more effectively than the other models. RandomForestRegressor is particularly strong in terms of its ability to generalize to unseen data, making it a promising model for this dataset.

Additionally, the insights gained from our study can serve as a valuable reference for researchers and practitioners selecting machine learning algorithms for similar datasets. Our findings guide choosing algorithms for classification tasks, especially when dealing with datasets that share characteristics similar to ours.

REFERENCE

[1] https://www.irjet.net/archives/V10/i3/IRJET-V10I342.pdf

[2] https://www.geeksforgeeks.org/ola-bike-ride-request-forecast-using-ml/

[3] https://www.projectpro.io/project-use-case/ola-bike-rides-request-demand-forecast