Machine Learning 1

GROUP ASSIGNMENT
ON
CAPITAL BIKESHARE

By - Pranjal Shukla, Saloni Sharma, and Yukti Handa

TABLE OF CONTENTS

1. Overview
2. Data Collection
3. Exploratory Analysis
4. Predictive Modeling
5. Performance Evaluation
6. Decision-making recommendations
7. Findings and Conclusions
8. Limitations

Overview

Data regarding rider behavior, including where they go, when they ride, and how far they go, is provided by Capital Bikeshare. The Capital Bikeshare Data Licensing Agreement governs the usage and analysis of this data by the general public. The journey duration, start and end times, station names and numbers, bike ID numbers, and member types are all part of the trip history data that may be downloaded. The Generic Bikeshare Feed Standard format is used to disseminate the system's real-time data. In order to evaluate usage patterns and growth strategies, Capital Bikeshare also conducts surveys of its membership, and its development plans every 2 years.

Our goal is to thoroughly analyze the Capital Bikeshare statistics in order to increase operational effectiveness. The lack of bicycle docks and bikes in regions with strong demand has been highlighted as the main issue. In the report, we will go into more detail about the two locations we chose to present as possible remedies to this problem. Our objective is to forecast the scenario and offer ideas for moving motorcycles to make it better.

Data Collection

For this project, we have used two datasets: Capital Bikeshare trip data and DC weather data. The Capital Bikeshare dataset consists of four different files, each of which contains a month's worth of trip data, covering the period from January 2022 to April 2022.

The DC weather dataset contains monthly summaries of weather-related parameters such as temperature, minimum and maximum temperature, precipitation, snowfall, and others.We have used this data to merge with the Capital Bikeshare dataset and to extract weather-related parameters for each trip. After merging the two datasets, The merged dataset contains 26 parameters, consisting of several weather-related parameters including temperature, windspeed, feels like, and wind direction,Out of these parameters Temperature, wind speed, feels like, and wind direction are the four weather-related variables we have chosen to further examine for this project.

Exploratory Analysis

The business understanding of Washington, D.C.'s Capital Bikeshare and the difficulties it encounters in maintaining client satisfaction. The success of the business depends on making sure that consumers can rent motorcycles and return them to docks. The team intends to perform a comprehensive analysis of the data to choose the best locations for the service and modify the number of ports offered at each stop in order to resolve these problems. This entails determining locations that need to be expanded by looking at how many vacant piers are present at each station. In order to predict the adjustments required to increase productivity, the team will use data from two stations. Repositioning the bikes can also help accomplish bike and dock availability.

The statistics used for the analysis are discussed by the members of the group. The data was taken from the Capital Bikeshare website and includes journey information from January to April 2022, including start and finish locations, time, date, coordinates, and member classification. In addition to this dataset, the team also used DC meteorological data to illustrate how weather conditions affect Capital Bikeshare utilization. The real highest and lowest temperatures, the sort, likelihood, and coverage of the precipitation, the humidity, and the dew factor are all included in the weather statistics. The team wants to use this data to understand how weather affects the access and use of the bikes, as well as to find trends in bike utilization.

Modeling

In order to forecast future events based on past and present data, predictive modeling uses data mining and machine learning. It enables businesses and organizations to make well-informed choices regarding future occurrences like sales, consumer behavior, and market trends. Organizations can predict and plan for future events and outcomes by using predictive modeling to find patterns and trends in data.

In this project, to predict the demand for bike sharing, various regression models were used throughout the modeling approach. The dependent variable is the quantity of bicycles dropped off each day, and the independent variables are different weather-related factors. In order to find the best model for forecasting demand and managing supply, we will then evaluate the model's performance and fine-tune the hyperparameters.We have used five different modeling approaches.

- **Linear Regression** - The statistical method of linear regression involves fitting a linear equation to the observed data in order to model the connection between a dependent variable and one or more independent variables. The aim of linear regression is to find the best-fitting line through the data which reduces the sum of the squared residuals between the predicted and actual values.
- **Ridge Regression** - Ridge regression is a regularized regression method that reduces overfitting by adding a penalty term to the cost function. The penalty term assists in getting the coefficients closer to zero and reduces the variance of the model as it is inversely proportional to the sum of the squared coefficients.
- **LASSO** - Least Absolute Shrinkage and Selection Operator (Lasso) is a method of regression that reduces the size of the coefficients, making some of them exactly zero. By doing this, the model's complexity is decreased and only the most crucial variables are chosen.
- **Elastic Net** - Elastic Net is a regularized regression method that combines the Ridge and Lasso regression approaches. It is utilized to get around the limitations of Lasso and Ridge regression, which can either pick variables or reduce the coefficients to zero. Elastic Net accomplishes this by adding both L1 and L2 regularization penalties to the cost function, which helps in the simultaneous selection of variables and shrinking of coefficients.
- **KNN** - K-Nearest Neighbors (KNN) is a machine-learning approach for classification and regression that is non-parametric. In order to generate a forecast based on the majority class or average value of the k-nearest neighbors, it identifies the k-nearest points in the training data to a new data point.

Performance Evaluation

Predicting pick-up and drop-off numbers in advance is made easier with the help of these Modelings. The transfer of bikes can be planned for later days using weather predictions. Bike utilization is greatly influenced by factors like weather, wind speed, and snow depth. These models are useful in creating an operational plan because they take all variables into account when making forecasts.

Additionally, Lasso is the most efficient of these models since it takes into account all of the major influencing factors, avoids redundancy of independent variables, and offers the pick-up and drop-off predictions with the lowest MSE. It prevents a recurrence of independent factors and is therefore the most effective model compared to the others.

I now realize how important it is to have a trustworthy example. We can schedule our working shift far in advance if the model's forecasts can be trusted, which is especially useful when managing hundreds of stations. Predictive data being available beforehand can therefore help in formulating an effective strategy. Overall, reliable prediction models can assist in enhancing the effectiveness and performance of bike-sharing systems.

21st & I St NW Pickups

	Model	MSE	Hyperparameters
0	Linear Regression	195.406000	N/A
1	KNN	237.696181	k = 4
2	Lasso CV	171.915000	0.869749
3	Ridge CV	167.542000	15.199111
4	Elastic Net	168.692000	0.173826

In this table, Ridge CV has the lowest MSE of 167.542000 which indicates the model with the best performance in terms of prediction accuracy. Elastic Net has the second lowest value of MSE of 168.692000, then comes the Lasso CV with 171.915000. The Mean squared error is a measure used to evaluate regression models and represents the average squared difference between the predicted and actual values. Hence, Ridge CV is the best performing model among all the five models. The coefficients of Ridge CV model are mentioned below -

```
The coefficients are:
temp                       0.728630
feelslike                  0.739190
windspeed                 -0.197804
winddir                    0.022961
icon_partly-cloudy-day     0.440683
icon_rain                 -0.050630
icon_snow                 -0.509394
icon_wind                  0.000000
dtype: float64
```

21st & I St NW Dropoffs

	Model	MSE	Hyperparameters
0	Linear Regression	175.432000	N/A
1	KNN	210.340495	k = 8
2	Lasso CV	141.159000	1.047616
3	Ridge CV	138.243000	21.049041
4	Elastic Net	137.014000	0.45468

In this table, Elastic Net has the lowest MSE of 137.014000 indicating that the model is better at predicting the outcome variable and has a better fit to the data. Ridge CV has the second lowest MSE of 138.243000 followed by Lasso CV with 141.159000. The Mean squared error is a measure used to evaluate regression models and represents the average squared difference between the predicted and actual values. Hence, Elastic Net is the best performing model among all the five models. The coefficients of Elastic Net model are mentioned below -

```
The coefficients are:
temp                        5.440714
feelslike                   4.714794
windspeed                  -2.005697
winddir                     0.030760
icon_partly-cloudy-day      2.932530
icon_rain                  -1.037897
icon_snow                  -1.855689
icon_wind                   0.000000
dtype: float64
```

21st St & Pennsylvania Ave NW Pickups

	Model	MSE	Hyperparameters
0	Linear Regression	39.730000	N/A
1	KNN	35.799911	k = 7
2	Lasso CV	38.874000	0.148497
3	Ridge CV	37.150000	12.915497
4	Elastic Net	38.715000	0.113269

In this table, KNN Regressor has the lowest MSE of 35.799911 indicating that the model is better at predicting the outcome variable and has a better fit to the data. Ridge CV has the second lowest MSE of 37.150000 followed by Elastic Net with 38.715000. The Mean squared error is a measure used to evaluate regression models and represents the average squared difference between the predicted and actual values. Hence, KNN Regressor is the best performing model among all the five models. The coefficients of KNN Regressor model are mentioned below -

```
In [38]: # optimal k
         np.argmin(mse_test)+1

Out[38]: 7

In [39]: # optimal MSE
         min(mse_test)

Out[39]: 35.799911268855375
```

21st St & Pennsylvania Ave NW Drop offs

	Model	MSE	Hyperparameters
0	Linear Regression	46.2330	N/A
1	KNN	61.5825	k = 5
2	Lasso CV	51.3540	0.453488
3	Ridge CV	51.7610	24.770764
4	Elastic Net	51.7060	0.436081

In this table, Linear Regression has the lowest MSE of 46.2330 indicating that the model is better at predicting the outcome variable and has a better fit to the data. Lasso CV, Ridge CV and Elastic Net have very similar MSE values of 51.3540, 51.7610 and 51.7060. However, Lasso CV has the second lowest MSE followed by Elastic Net. The Mean squared error is a measure used to evaluate regression models and represents the average squared difference between the predicted and actual values. Hence, Linear Regression is the best performing model among all the five models. The coefficients of Linear Regression model are mentioned below -

```
temp                       0.174171
feelslike                  0.253217
windspeed                 -0.282004
winddir                    0.001496
icon_partly-cloudy-day     2.487210
icon_rain                 -2.624711
icon_snow                 -0.855634
icon_wind                  0.993134
dtype: float64
```

Key Insights

The Lasso model is the best at predicting exactly when and where bikes will be picked up and dropped off. This implies that the bikeshare system can use the Lasso model to develop a reasonable forecasting system that can support station planning, scheduling, and operational decision-making.

To continuously improve the prediction models and maximize bike use, the system can implement performance evaluation measures like MSE. The system can boost the accuracy of the forecasts and the effectiveness of bike sharing operations by constantly assessing and improving the models.

The system can forecast when bikes will be picked up and dropped off at the 21st St & Pennsylvania Ave NW station using the KNN Regressor model. For this station, this model has the lowest MSE, indicating that it is the most accurate for making predictions.

For forecasting bike drop offs at the 21st St & Pennsylvania Ave NW stop, the best model is linear regression. To ensure that bikes are always accessible for usage by riders, this model can be used to optimize the station's bike allocation.

Findings and Conclusion

Scenario 1				
	21st street		**Pennsylvania**	
	Pickups	Drop Offs	Pickups	Drop Offs
Predict	36	39	23	13
Actual	47	45	15	1

Scenario 2				
	21st street		**Pennsylvania**	
	Pickups	Drop Offs	Pickups	Drop Offs
Predict	23	22	17	8
Actual	16	10	13	13

1. Compared to 21st St & Pennsylvania Ave NW station, 21st & I St NW station is busier. This indicates that more bikes can be sent by the system to the 21st & I St NW station to accommodate the increased demand. To meet the increased demand, the system can also consider increasing the capacity of the 21st & I St NW station.
2. Two Capital Bikeshare stations at 21st & I St NW and 21st St & Pennsylvania Ave both have 53 docks that are operable. Notably, the station at 21st St & Pennsylvania Ave has 18 docks, compared to 35 docks at the station at 21st & I St NW. (reference -https://secure.capitalbikeshare.com/map/)
3. In scenario 1, although there were 75 pickups and drop offs at the Capital Bikeshare station at 21st and I St NW in a single day, there are only 35 docks at this station. Similar to the 21st St & Pennsylvania Ave station, which has only 18 docks, there were 36 pickups and drops off per day there. Nonetheless, at this time, we are unable to precisely identify when the demand for bikes is highest, whether it is during the day, night, afternoon, or evening.

4. In scenario 2, we have seen situations where there are more pickups and drop offs than there are docks available at both stations, similar to scenario 1. We suggest that the overcrowded 21st & I St NW station receive at least 30 additional docks, and the overcrowded 21st St & Pennsylvania Ave station get at least 18 additional docks. This would help prevent issues with dock availability and guarantee that consumers have access to the bikes they require.

Limitations

Our current machine learning (ML) model has some limitations that should be taken into consideration when interpreting the results. Firstly, the model is based solely on the daily number of pickups and drop-offs at each Capital Bikeshare station, and does not take into account the specific time zones or rush hours when demand for bikes may be higher. As such, we cannot be certain about peak demand or potential traffic at each station, which could impact our recommendations for dock allocation.

Secondly, the model only allows us to swap and exchange between two docks at a time. This means that we were not able to consider adding or removing larger numbers of docks, which could significantly impact the performance of the bike-sharing system.

Therefore, it is important to keep these limitations in mind when evaluating the recommendations generated by our ML model, and to consider additional factors such as user feedback and real-time usage data in order to optimize the allocation of bike docks and improve the overall performance of the system.