Machine Learning —
Capital Bikeshare Project

Table of Contents

Executive Summary

We are assigned to run a detailed analysis on the Capital Bikeshare data that can help achieve operational excellence. The main issue highlighted is the lack of supply for both docks and bicycles in areas with high demand. We proceed forward with two stations to provide possible solutions to this problem and further explain it in the report. The objective is to predict and provide recommendations for the potential repositioning of bikes.

Business Understanding

Capital Bikeshare is a bicycle-sharing system, situated in Washington D.C. Over the course of the last couple of years, there has been a significant increase in the usage of this service. However, the business is constantly struggling to ensure that customer satisfaction remains their top priority. Logistically, the business model runs on two simple needs; there are bikes available nearby whenever a customer wants to ride one and similarly, the docks are available when they want to return it. We are aiming to solve it by running a holistic analysis of the data provided in order to reach a fair conclusion.

A strategic decision that needs to be delivered in the course of this analysis is to determine where to offer the service. It is essential that the supply of bikes and docks remains ample where there is a significant demand for it, in order to reduce cost and keep customer satisfaction above par. Another tactical decision that we aim to achieve with the analysis is to adjust the number of docks available at each station. This requires an analysis based on how many empty docks are available at each station as well as which stations might require expansion, based on them being at capacity. The operational challenge, as aforementioned, of bike and dock availability can be achieved by repositioning the bikes. We proceeded forward with two stations that each acted as pick up and drop off, providing us with 4 data points in total. This in turn will provide us with estimated values of changes that have to be in place in order to achieve efficiency.

Data Understanding

The dataset we used was accessed from the official Capital Bikeshare website. For convenience purposes, we used data from 202201 to 202204 which included the trip details, including start/pick up station, time, date, coordinates and the member type. The same information was available for the end/ drop off station as well. Apart from this dataset, we used the DC weather data which included date, actual maximum and minimum temperatures as well as what it felt like, precipitation type, probability and coverage, humidity and dew factor. The reason for using this weather data is to visualize the impact the weather conditions have on the usage of capital bike share.

Modeling

During the modeling process, various regression models were utilized to predict the demand for sharing bicycles. The target variable is daily drop offs and the independent variables are all kinds of weather features. Subsequently, we will evaluate the model's performance and adjust the hyperparameters to identify the optimal model for predicting demand and managing availability.

a. Linear regression: We just regress the demand on 26 weather features. The main advantage of linear regression is its ease of interpretation. The coefficients in the linear equation provide a straightforward measure of the strength and direction of the relationship between the dependent variable and each independent variable. This makes it easy to understand how each independent variable affects the dependent variable, which can help in making informed decisions.
b. Ridge regression: In the Ridge regression, we used regularization to reduce the impact of multicollinearity and overfitting problem. The L2 norm penalty would constrain the sum of the squared values of the regression coefficients. The penalty term encourages the model to shrink the coefficients towards extremely small, very close to zero but not exactly zero. By doing so, Ridge regression can effectively reduce the variance of the model.Eventually, the MSE in the test dataset could reduce.
c. LASSO: LASSO regression is a type of linear regression that uses regularization to prevent overfitting and improve model performance. Because of the L1 norm penalty, it encourages the model to shrink the coefficients of less important features towards zero, effectively performing feature selection by identifying and excluding irrelevant or redundant features. In our analysis, Lasso regression is particularly useful. Given many similar weather features and some of them did not have the predictive power, LASSO can shrink MSE.
d. Elastic Net: Elastic Net is another regularized regression method that combines the strengths of Lasso and Ridge regression. By adding both L1 and L2 penalties to the objective function, Elastic Net is able to simultaneously address the problems of multicollinearity and overfitting in high-dimensional datasets. Elastic Net can not only provide feature selections but also reduce the variance in the model. Therefore, Elastic Net can handle correlated predictors and avoid overfitting better than Lasso and Ridge regression alone.
e. KNN regressor: K-Nearest Neighbors (KNN) is a non-parametric machine learning algorithm that can be used in regression problems. KNN works by finding the k-nearest neighbors of a given data point in the feature space and predicting the output variable based on the average or median of the target values of those neighbors. The merit of KNN is to explain non-linear relationships rather than linear regression. This makes KNN regression a more flexible and adaptable model than linear regression alone.

Evaluation

We tune the model to maximize the model performance that is why hyperparameter tuning is applied in a linear regression model to reduce the variance error and avoid overfitting. Similarly for KNN models choosing the optimal K value is important that minimizes the MSE. This is why we split the data into three sets: training, validation and test sets.

We can use L1(Lasso) regularization when our model contains many useless variables. As this is the case where Lasso works best. Similarly, L2(Ridge regression) works best when most of the variables in the model are useful. Ridge regression will help reduce the variance in our model by shrinking the parameters by making them less sensitive but not removing them from the model. Elastic regression is a combination of the two methods above. In elastic regression we have two lambdas one for lasso and the other for ridge regression.

If lambda 1 > 0 and lambda 2 = 0. - we get Lasso
If lambda 1 = 0 and lambda 2 > 0. - we get ridge
If lambda 1 > 0 and lambda 2 > 0. - we get a hybrid elastic net regression

Elastic regression is best used when there is a correlation between the parameters/variables in our model. Lambda values can range from any value between zero to infinity. If the lambda value is equal to zero hence, the penalty is also then equal to zero as it nullifies the effect. As the value of lambda increases the slope gets smaller. We use cross validation to decide which value of lambda is best to use that will result in lowest variance in our model. Hyperparameter tuning allows us to control the learning process. The hyperparameters that are selected then can help improve the learning of the model. Thus hyperparameter tuning is very advantageous as it can help increase the accuracy of a machine learning model by running multiple trials.

Cross validation allows us to compare different machine learning algorithms and get an idea of how well they will practice and work for our given dataset. Cross validation is a resampling method or technique that uses different sets of the data to test and train a model on different iterations.

21st & I St NW Pickup model evaluation

	Model	MSE	Hyperparameters
0	Linear Regression	213.572	N/A
1	KNN CV	195.380	k = 2
2	Lasso CV	143.928	0.792483
3	Ridge CV	147.980	65.793322
4	Elastic Net CV	144.573	0.844154

21st St & Pennsylvania Ave NW Pickup model evaluation

	Model	MSE	Hyperparameters
0	Linear Regression	47.308000	N/A
1	KNN CV	39.464568	k = 6
2	Lasso CV	26.123000	0.792483
3	Ridge CV	31.032000	65.793322
4	Elastic Net CV	28.391000	0.84239

21st & I St NW Drop offs model evaluation

	Model	MSE	Hyperparameters
0	Linear Regression	233.961000	N/A
1	KNN CV	208.150864	k = 3
2	Lasso CV	133.142000	0.722081
3	Ridge CV	128.597000	65.793322
4	Elastic Net CV	129.268000	0.80005

21st St & Pennsylvania Ave NW Drop offs model evaluation

	Model	MSE	Hyperparameters
0	Linear Regression	55.297000	N/A
1	KNN CV	67.913889	k = 2
2	Lasso CV	44.019000	1.047616
3	Ridge CV	45.020000	91.116276
4	Elastic Net CV	44.231000	1.041654

Note: CV is Cross-Validation in the model, K=5. Hyperparameters of the Lasso, Ridge, and Elastic Net model refer to Best Alpha.

The above four tables show the results of the various models for the two stations based on different scenarios: pick-up and drop offs.

Overall, looking at the results altogether we can see that the linear model has the highest MSE amongst all the different models except for the 21st Street & Pennsylvania Ave NW dropoff where KNN model has the highest MSE. That is then followed by the KNN model which has the second highest MSE overall except for the 21st street & Pennsylvania Ave NW dropoff where the Linear model has the second highest MSE. In general the Lasso model is the best model altogether as it has the lowest MSE except for the 21st street & I street NW drop off where the Ridge model has the lowest MSE.

Because of the smallest MSE in the test dataset, we chose Lasso regression to develop the prediction model. In the cross-validation process, we received the best alpha to optimize the model. The hyperparameters tuning can refer to the above table: the best alpha for the Lasso CV model is shown.

From the Lasso regression output, we can see that many coefficients were shrinkage to 0. The feature selection encourages the prediction performance on the test dataset.

21st & I St NW Pickup Lasso model	**21st & I St NW Drop-off Lasso model**	**21st St & Pennsylvania Ave NW Pickup Lasso model**	**21st St & Pennsylvania Ave NW Drop-off Lasso model**
The coefficients are: tempmax 2.061753 tempmin 0.000000 temp 6.690140 feelslikemax 0.000000 feelslikemin 0.000000 feelslike 0.000000 dew 0.000000 humidity -2.245071 precip -0.000000 precipprob -0.000000 precipcover -2.495638 snow 0.000000 snowdepth -3.228062 windspeed -2.590788 winddir 0.481611 sealevelpressure 0.000000 cloudcover 0.000000 visibility 0.790024 solarradiation -0.000000 solarenergy 2.318058 uvindex 0.000000 moonphase 0.765663 icon_partly-cloudy-day 0.000000 icon_rain -0.000000 icon_snow 0.000000 icon_wind 0.000000	The coefficients are: tempmax 0.000000 tempmin 0.000000 temp 9.629816 feelslikemax 0.000000 feelslikemin 0.000000 feelslike 0.000000 dew 0.000000 humidity -1.198794 precip 0.000000 precipprob -0.000000 precipcover -1.312996 snow -0.000000 snowdepth -3.974668 windspeed -3.051639 winddir -0.000000 sealevelpressure 0.517770 cloudcover 0.000000 visibility 1.661285 solarradiation -0.000000 solarenergy 2.742897 uvindex 0.000000 moonphase 2.364949 icon_partly-cloudy-day 0.000000 icon_rain -0.000000 icon_snow -0.051349 icon_wind 0.000000	The coefficients are: tempmax 1.439375 tempmin 0.000000 temp 0.000000 feelslikemax 0.000000 feelslikemin 0.000000 feelslike 2.410625 dew 0.000000 humidity -0.000000 precip -0.000000 precipprob -0.000000 precipcover -0.000000 snow -0.000000 snowdepth -0.000000 windspeed -0.611661 winddir 0.000000 sealevelpressure -0.000000 cloudcover 0.000000 visibility -0.000000 solarradiation 0.000000 solarenergy 2.061292 uvindex 0.869532 moonphase 0.704540 icon_partly-cloudy-day 0.037787 icon_rain -0.000000 icon_snow -0.000000 icon_wind 0.000000	The coefficients are: tempmax 0.000000 tempmin 0.000000 temp 0.000000 feelslikemax 0.000000 feelslikemin 0.327060 feelslike 3.599647 dew 0.000000 humidity -0.000000 precip 0.000000 precipprob -0.000000 precipcover -0.000000 snow -0.000000 snowdepth -0.807590 windspeed -0.205237 winddir -0.000000 sealevelpressure -0.000000 cloudcover 0.000000 visibility -0.000000 solarradiation 0.000000 solarenergy 3.418568 uvindex 0.000000 moonphase 0.000000 icon_partly-cloudy-day 0.000000 icon_rain -0.000000 icon_snow -0.000000 icon_wind 0.000000

Discussion and Conclusion

We applied the model to predict the second and the third test data point, annotated as scenario 1 and 2:

	Scenario 1			
	21st & I St NW		**21st St & Pennsylvania Ave NW**	
	Pickups	Drop-Offs	Pickups	Drop-Offs
Predict	34.35	37.12	24.38	13.41
Actual	47	45	15	1

	Scenario 2			
	21st & I St NW		**21st St & Pennsylvania Ave NW**	
	Pickups	Drop-Offs	Pickups	Drop-Offs
Predict	22.01	20.86	16.28	8.45
Actual	16	10	13	13

1. Regardless of the scenario, it seems that 21st & I St NW is expected to have higher demand than 21st St & Pennsylvania Ave NW: Pickups and Drop-offs are expectedly higher. That is, 21st & I St NW is busier than 21st St & Pennsylvania Ave NW.
2. From the existing situation, 21st & I St NW can deploy a total 16 bikes or 16 docks if empty, while Pennsylvania Ave NW has 19 spots. Combining the capacity, 35 spots are flexible.
3. **Scenario one**: we can see the daily pickup and drop offs for 21st & I St NW are 34 and 37 respectively. However, we only have 16 spots. To fulfill the demand, we need to expand the capacity to 71 spots (deployed 34 bikes and 37 docks). That could satisfy the daily needs but also have the idle capacity issue. We assumed the daily demand comes up with four periods: morning, noon, afternoon, and night. The demand distributes evenly. With that said, in each period, the pickup number should be 8.5 (34/4) and 9.25 (37/4) drop offs on 21st & I St NW; 6 pickups and 3.35 drop offs from 21st St & Pennsylvania Ave NW. Under this assumption, we recommend deploying 9 bikes at 21st & I St NW and 6 bikes at 21st St & Pennsylvania Ave NW; the rest docks are empty. If users cannot find a dock on 21st & I St NW, they can visit 21st St & Pennsylvania Ave NW station to drop off the bike.
4. **Scenario two**: if we hold the same 4 traffic periods assumption, the demands in each period are: 21st & I St NW 5.5 pickups and 4 drop-offs, and 21st St & Pennsylvania Ave NW 4 pickups and 2 drop-offs. We recommend the Capital Bikeshare deploy accordingly. As for the extra capacity, they can consider the other stations' demand and thus decide how to leverage the capacity.

Limitations

Our model predicts the expected demand for the whole day, so it's hard to manage bicycles regarding the minutes or different traffic periods. The refined model can consider more nuanced situations, such as the pickup's arrival rate within a day. By doing so, the prediction can fit into real-time needs. Moreover, the weather is the main independent variable in our regression model. This assumption would not hold, if there is a time trend or seasonality, such as a travel season in the summer. To refine the model, more delicate predictors can be involved, e.g. time series. In the end, our model only considers the expected value of bike demand. When it comes to inventory management, the standard deviation is important to predict dynamic demand. Also, the customer satisfaction rate is important to manage inventory volume. With more information, Capital Bikeshare can optimize its overall performance.