DECISION 520Q - Data Science for Business

Section 101B Team 05: Chuyou Chen, Sidney Pepple, Meng Ru, Vivian Wang, Yiqing Wang, Junyi Zhang, Xinyi Zhou

Capital Bikeshare

1. Business Understanding

1.1 Research Background

The uniqueness of the bike-share industry is the primary reason that we chose this project. It is rapidly becoming popular as a primary use of transportation, especially in urban areas and campuses. In the past, people cycled primarily as a form of sport or exercise, but now, households and individuals make use of bicycles to transport themselves.

Due to the pandemic, the bike-share industry experienced a decline in the number of bike-share customers, leading to a huge dip in revenue, but it is now gradually recovering as cities and businesses reopen. Research on bike and scooter rental markets predicts that the revenue for the bike-share industry will increase from $2.5 billion in 2019 to $10.1 billion in 2027, which is due to people's need for easy commuting and increasing use in the food delivery industry to avoid traffic[1].

1.2 Business Value

Compared to traditional public transportation, bike sharing can provide people with a new way to travel and even adapt people to a more environmentally friendly way of living. However, bearing the high initial capital expenditure and ongoing depreciation expense, the bike sharing companies are facing risks of loss and a problem on **how to maximize allocation efficiency to gain profits**.

To be a successful player in the bike-sharing system, a company needs to understand what the future trend of the industry is and to balance its supply and customers' demand accordingly. If the company oversupplies, the bikes not being used become a waste of resources. If the company undersupplies, it will not be able to meet all the demand, and thus lose a portion of potential revenue. Our research will **help companies to understand the demand for shared bikes** among the people, especially in Washington D.C., and **then predict the optimal number of supplies based on this demand in order to maximize profit.**

2. Data Understanding

2.1 Data Source

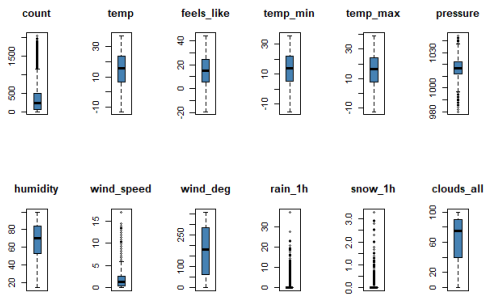
We believe that the demand for bike share should be closely related to weather conditions and people's need for commuting. We found the 2018 Jan - 2021 Aug bike share trip data in Washington D.C. from Capital Bikeshare[2]. Each row stands for one unique bike share rental record. In addition, we found the weather information for the same period from Open Weather Map[3]. Each row describes the weather information, including temperature, wind speed, pressure, etc., for each hour. Finally, we found the holiday schedule for each year from DC.gov[4] and calculated which day is a working day (Mon – Fri). Then we counted the number of bike shares during each hour, and inner join with the hourly weather condition, whether the day is a working day or a holiday. The first 5 rows of the data are shown below:

datetime	count	holiday	workingday	temp	feels_like	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	rain_1h	snow_1h	clouds_all	weather_main
1/1/2018 0:00	34	1	0	-7.17	-12.73	-8.56	-7.09	1030	53	3.6	310			20	Clouds
1/1/2018 1:00	49	1	0	-7.35	-13.81	-9.03	-7.15	1030	49	4.6	310			1	Clear
1/1/2018 2:00	37	1	0	-7.88	-14.05	-9.03	-7.69	1031	52	4.1	310			1	Clear
1/1/2018 3:00	9	1	0	-8.1	-14.32	-9.36	-7.89	1031	49	4.1	310			1	Clear
1/1/2018 4:00	12	1	0	-8.19	-14.43	-9.46	-8.09	1031	49	4.1	330			1	Clear

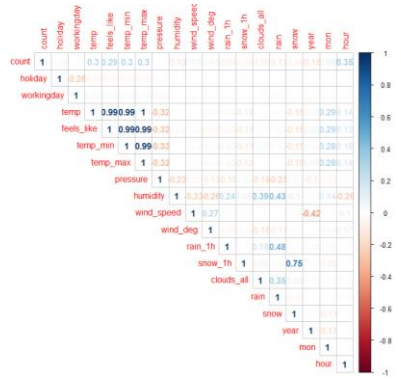
There exist some potential biases in our data. Firstly, Capital Bikeshare does not provide the data for 2020 April on its website, potentially due to Covid-19. Missing data means loss of information which might reduce the statistical significance of some variables in the model. We also realize that the customer

demand might be underestimated by the count of bike share trips when the bikes are undersupplied and some of the demand is not reflected in the trip data.

2.2 Data Visualization



To better understand the distribution of data, we created boxplots for numeric variables. The count of rentals is right skewed with quite a few outliers of high demand. Temperature, "feels like" temperature, minimum temperature and maximum temperature follow similar distributions, but "feels like" temperature has a wider range. Variables including pressure, wind-speed, rain_1h and snow_1h have extreme values which may cause bias in the results and degrade the efficiency of the data. Those outliers may affect the process of predicting process and result in overestimated or underestimated counts.



Besides, we used a correlation matrix to find the correlation between variables. From the matrix, we find that temperature, "feels like" temperature, min temperature, max temperature, humidity and rain are more influential than other variables on the count of rentals. In addition, temperature, "feels like" temperature, min temperature and max temperature are highly correlated with each other which might reduce the precision of the estimated coefficients and need to be addressed when building models.

We are interested in how people perceive temperature differently from the actual temperature. To achieve this, we divided the temperature into 5 groups and plotted line charts for temperature and "feels like" temperature for each group. "Feels like" temperature (orange line) is always lower than actual temperature (blue line) and the gap is more obvious when temperature is lower than 0 or higher than 30. While the "feels like" temperature is calculated by some meteorologists based on actual temperature, humidity, wind speed, etc., we believe that people might be even more sensitive to extreme weather and therefore lead to a wider gap not captured by the model.

From the graph on the top-left corner, we can see the highest demand for bike-share appears when the temperature is between 20 to 30°C. We can know from the bottom graph that from April to October, the average temperature is between 20 to 30°C, and the number of counts is also relatively higher than the rest of the year. This acknowledges our findings based on the top graphs. In conclusion, bike share demand will be higher in temperatures between 20-30°C.

3. Data Preparation

To prepare the data for further modeling, we cleaned the dataset and modified some data types.

Firstly, we examined if there were any missing values. All the missing values appeared in “rain_1h” and “snow_1h” columns. The values are missing because it was not raining or snowing during the last hour and therefore, the precipitation data was not collected. To deal with the missing values, we created two dummy columns “rain” and “snow” to indicate whether the data is missing or not, which also stands for whether it was raining or snowing during the last hour, and then impute 0 to all missing values in columns “rain_1h” and “snow_1h”.

Next, since the “datetime” column is in a string format and is hard to use in the following modeling, we divided the column into “year”, “mon” (month), and “hour”, and remove the original “datetime”. We then modified the three variables and other variables including “holiday”, “workingday”, “rain”, “snow”, “weather_main” into factor columns because they are categorical instead of numeric by nature.

A detailed description of our final dataset is shown in Appendix 2.

4. Modeling

4.1 Identify the Critical Variable

We started our modeling process by using double selection method to figure out the critical variables in the dataset and see how “count” variable change in accordance with the change of independent variables. All variables except “temp_min” and “snow_1h” significantly affect the ‘count’ variable, suggesting that the variables chosen in our dataset are meaningful.

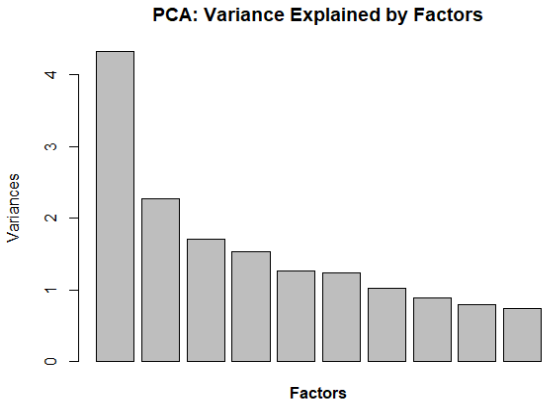
Holding other factors constant, raining and snowing can significantly affect the demand of the shared bike. Raining in the last one hour can decrease the demand by 83 and snowing in the last one hour decreases the demand by 133. (Notice that “snow_1h”, the precipitation of snow in the last one hour, is insignificant, but “snow”, whether it is snowing in the last one hour, has a large influence.) The average demand during

holiday is about 55 lower than that during non-holiday. Besides, since the coefficient of year is –63.61, we expect that the result is due to the lockdown during the pandemic.

We only discussed here some significant variables, and the full result can be found in Appendix 3.

4.2 Unsupervised Learning

To increase interpretability but at the same time minimize information loss, we continue to do exploratory analysis using primary component analysis method. Considering the limited input data types, we dropped "weather_main" and believed that the information is embedded in other variables.



The variance explained by factors is plotted in the left graph. According to the graph, the first four factors can explain most of the variance. Thus, in the following analysis, we mainly focus on these four factors. The correlation between factors and variables are as follows, rounded to two decimals.

```
> round(loadings, 2)
             PC1   PC2   PC3   PC4
holiday    -0.03 -0.02  0.08 -0.02
workingday  0.03  0.02 -0.05  0.02
temp        0.47  0.05 -0.08  0.00
feels_like  0.47  0.03 -0.09 -0.03
temp_min    0.47  0.04 -0.08  0.02
temp_max    0.47  0.05 -0.09 -0.02
pressure   -0.20  0.21 -0.02 -0.14
humidity    0.03 -0.52  0.02 -0.08
wind_speed  0.00  0.17  0.12  0.65
wind_deg   -0.02  0.24  0.05  0.30
rain_1h     0.05 -0.36  0.17  0.16
snow_1h    -0.10 -0.13 -0.65  0.17
clouds_all  0.02 -0.41  0.02  0.16
rain        0.04 -0.48  0.22  0.18
snow       -0.12 -0.14 -0.64  0.18
year        0.02 -0.07 -0.14 -0.51
mon         0.17 -0.01  0.09  0.07
hour        0.08  0.12 -0.06  0.22
```

It shows that PC1 is positively correlated with "temp", "feels_like", "temp_min", "temp_max", which means that PC1 mainly depicts times with high actual temperature and body temperature. PC2 is negatively correlated with "humidity", "rain_1h", "clouds_all", "rain" and is positively correlated with "wind_deg", which means that PC2 mainly depict dry and windy times with low cloud cover and without rain at least in the past one hour. PC3 is negatively correlated with "snow_1h", "snow" and is positively correlated with "holiday", which means that PC3 mainly depicts times without snow and is likely to be holidays. PC4 is positively correlated with

"wind_speeed", "wind_deg", "rain_1h" and is negatively correlated with "pressure" and "year", which means that PC4 depict windy times with low air pressure and is likely to have rained in last one hour. And PC4 mainly depicts times in earlier years.

```
> round(colMeans(cati1),2)
      PC1       PC2       PC3       PC4 count_cati
    -0.65     -0.39      0.00     -0.24      1.00
> round(colMeans(cati2),2)
      PC1       PC2       PC3       PC4 count_cati
    -0.39     -0.16      0.07      0.08      2.00
> round(colMeans(cati3),2)
      PC1       PC2       PC3       PC4 count_cati
     0.11      0.07      0.01     -0.03      3.00
> round(colMeans(cati4),2)
      PC1       PC2       PC3       PC4 count_cati
     0.93      0.48     -0.09      0.19      4.00
```

To better understand how PCs explain 'count', we stratify the column using 1st quantile, median and 3rd quantile and calculate the column mean for each group. Based on the result, it shows that shared bikes are usually popular at hot and dry times, which aligns our findings in visualizations.

4.3 Predictive Models

After unsupervised learning, we started to build predictive models. Since our dependent variable is the numeric "count", we mainly consider two types of models: linear regression models and tree models.

The linear regression model is the simplest equation, the modeling speed is fast, and the model can take different interactions between variables. Besides, the coefficient from linear regression model is easy to interpret and can provide us insights about the influence of each variable. The cons are the model might be overly simplistic and we are assuming a linear relationship between variables.

We also applied Lasso and Ridge to linear regression model to control overfitting. Ridge controls by shrinking some of the coefficients and Lasso controls even further by forcing some to zero. However, the interpretability is comparatively low, and the selected features might bring in bias. Besides, since Lasso reduces some coefficients to zero, and only selects one feature from a group of correlated features, it might eliminate some variables that might be interesting to look at.

The regression tree model accepts different types of inputs and can handle non-linear relationships between variables. It is also easy to interpret by reading the plot. However, the single tree model is unstable

and is highly influenced by the split. Thus, we applied random forest to average the results across multiple trees to control for overfitting. However, we had to limit the complexity of the random forest (ntree=20) to control the training time and it is hard for us to interpret each tree in the model.

By modelling demand of the market using data from 2018-2021, companies can understand how exactly the demands vary with unique features. As a result, companies can estimate the demand for shared bikes reasonably. It can then accordingly develop information and operation system to optimize bicycle allocation, adjusting the quantity supplied to meet customers' demand and expectations.

This analysis provides the company with a basic understanding of the dynamics of the market, which is the basis of success. More detailed analysis can be conducted if we are provided with more information.

5. Evaluation

To better measure the performance of each model, we randomly selected 30% of our observations in the training set as holdout sample. We then calculated out of sample R^2 (OOS R^2) to compare the performance of each model (Appendix 4).

Firstly, we run a linear model and use step() function to control overfitting. The OOS R^2 is 63.05%. We tried multiple interactions and found interacting all variables with "Workingday" providing us with the highest OOS R^2 73.89%. After that, we applied Lasso, Post Lasso, and Ridge models. The highest OOS R^2 was achieved by Post Lasso estimates, with OOS R^2 of 73.89%. For the tree models, the single tree model did not perform well, but the random forest model improved our OOS R^2 to **86.69%**.

By comparing OOS R^2 and plotting the actual count and our predicted count, we found that linear regression model might not capture the relationship as well as trees, and thus we decided to use random forest as our final model. We then applied K-fold cross validation to random forest model and calculated the OOS R^2 for each fold. The average performance is 87.25%. The 1st quartile is around 86.84% and the

3rd quartile is around 87.56%. The random forest model is not only performing much better than the other models in this scenario, but also exhibiting a very stable performance across different folds.

After forecasting the market demand for the hour, our next step is to maximize expected profit by finding the appropriate supply. We provide a basic model below to find the appropriate supply level.

p= Price per trip *c = Operating cost to supply one bike*

s= Number of bikes supplied *d = Number of bikes demanded*

E[Profit] = E[Revenue] - E[Cost] = E[p * min(s, d)] - c * s = p * E[min(s, d)] - c * s

= p * [s * P(s < d) + d * P(d < s)] - c * s

We made the following assumptions:

- Price: $1 per trip[5] (To simplify the calculation, we are not considering the fixed membership fee per month)
- Cost per bike per hour: $0.28 (Appendix 5) (We use the four-year average of operating costs divided by number of bikes in Arlington area)

After making the assumptions, we used the random forest model to run a larger random forest model based on the weather on Oct 7, 2021 18:00 to produce 500 different predictions of the potential demand level. By using for loop to run through different numbers of bikes that the company can supply, and using the distribution of 500 predictions (Appendix 6) to calculate the probability that supply will be more than demand, we can find the supply level that will maximize the profit. For this hour specifically, we predicted that by supplying 417 bikes, they can maximize their profit to $ 288.02.

6. Deployment

Ideally, the company can use the random forest model to find the potential demand for the hour and then estimate the optimal number of bikes supplied to maximize the profit. However, there might be some

difficulties in achieving the optimal supply. For example, it is difficult to monitor and regulate people's behaviors in the sharing industry. There is a high possibility that people pick up and drop their share bikes at different locations. Behaviors like this are hard to predict and control. Besides, to determine the supply, the company should also strike a balance between quantitative and qualitative factors in their decision process. Apart from profit, there might be other things that a company values more, such as brand image. For example, too much undersupply might make customers disappointed and hurt the brand image. If there is too much oversupply, the bike parking area might be quite messy and will hurt the brand image as well. Therefore, the company should also consider the trade-off between profit maximization and customer satisfaction maximization.

We try to minimize the ethical issues by using objective data in our model and not resampling the data repetitively to get better performance. However, there are some other potential risks associated with our proposed plan. Besides data bias, we also ignore member pricing and assume operating costs per bike are fixed when determining supply. To mitigate the risk, we are limiting the scope to only Washington D.C., and provide only a general suggestion of balancing oversupply and undersupply. If we have more detailed data, we will be able to build a more realistic model and generate a more accurate profit prediction.

Please find all references and appendices in a separate file.