Predicting Bike Rentals

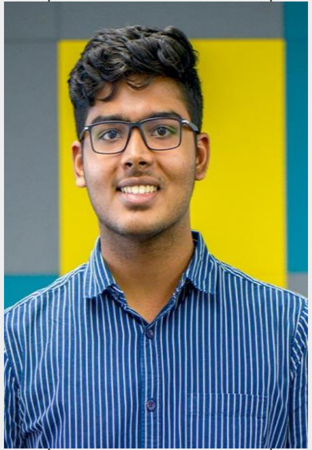
Introduction to Machine Learning Project

Group 10

MEET THE TEAM

Jenna Ferguson

Sankalp Kulkarni



Santhosh Ramkumar

Vishal Gupta

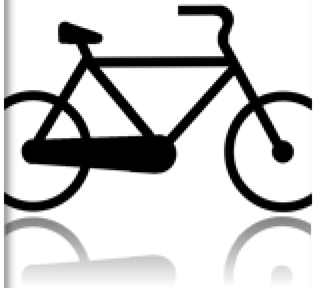
Meenal Gaba

Raghav Vaidya

Introduction and Objectives

Introduction

- There are **over 500 bike rental systems** across the world
- **Demand of bikes** on any particular day could vary depending upon many factors like:
  - Working day / Holiday
  - Time of the day
  - Weather on the day
  - ……..
- **Predicting the demand** for any bike rental service would help them better manage their resources and meet the demand more effectively



Objective

- **Objective is to predict the hourly demand** of bikes based on historical data of demand
- Our dataset contains hourly rental data from the **Capital Bikeshare program** in Washington DC, spanning a **2-year period**
- **We will test different models** like linear regression and Trees to determine the best possible predictions
- Based on the predictions we will lay down some **key recommendations** for the service provider

Deep-Dive into the Dataset

Data Snapshot

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
1/1/2011 0:00	1	0	0	1	9.84	14.395	81	0	3	13	16
1/1/2011 1:00	1	0	0	1	9.02	13.635	80	0	8	32	40
1/1/2011 2:00	1	0	0	1	9.02	13.635	80	0	5	27	32
1/1/2011 3:00	1	0	0	1	9.84	14.395	75	0	3	10	13
1/1/2011 4:00	1	0	0	1	9.84	14.395	75	0	0	1	1
1/1/2011 5:00	1	0	0	2	9.84	12.88	75	6.0032	0	1	1
1/1/2011 6:00	1	0	0	1	9.02	13.635	80	0	2	0	2
1/1/2011 7:00	1	0	0	1	8.2	12.88	86	0	1	2	3
1/1/2011 8:00	1	0	0	1	9.84	14.395	75	0	1	7	8
1/1/2011 9:00	1	0	0	1	13.12	17.425	76	0	8	6	14
1/1/2011 10:00	1	0	0	1	15.58	19.695	76	16.9979	12	24	36
1/1/2011 11:00	1	0	0	1	14.76	16.665	81	19.0012	26	30	56
1/1/2011 12:00	1	0	0	1	17.22	21.21	77	19.0012	29	55	84
1/1/2011 13:00	1	0	0	2	18.86	22.725	72	19.9995	47	47	94
1/1/2011 14:00	1	0	0	2	18.86	22.725	72	19.0012	35	71	106
1/1/2011 15:00	1	0	0	2	18.04	21.97	77	19.9995	40	70	110
1/1/2011 16:00	1	0	0	2	17.22	21.21	82	19.9995	41	52	93
1/1/2011 17:00	1	0	0	2	18.04	21.97	82	19.0012	15	52	67
1/1/2011 18:00	1	0	0	3	17.22	21.21	88	16.9979	9	26	35
1/1/2011 19:00	1	0	0	3	17.22	21.21	88	16.9979	6	31	37
1/1/2011 20:00	1	0	0	2	16.4	20.455	87	16.9979	11	25	36
1/1/2011 21:00	1	0	0	2	16.4	20.455	87	12.998	3	31	34
1/1/2011 22:00	1	0	0	2	16.4	20.455	94	15.0013	11	17	28
1/1/2011 23:00	1	0	0	2	18.86	22.725	88	19.9995	15	24	39

Note:

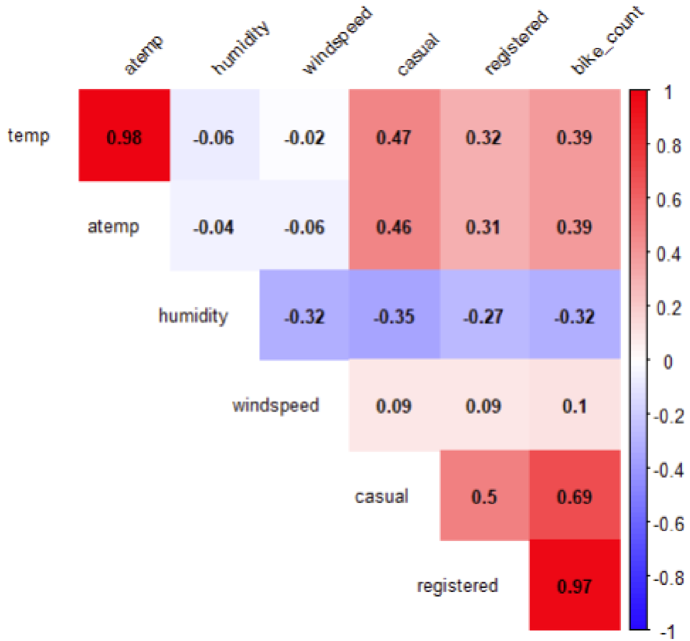
- ***Casual*** *and* ***registered*** *columns are a split of the* ***count*** *column depicting non-registered vs registered user rentals. Hence, we will likely not use them as input variables.*
- ***Datetime*** *will have to be split into four parts – Hour, Day, Month and Year*

Data Dictionary

Variable Name	Description
datetime	hourly date + timestamp
season	1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday	whether the day is considered a holiday
workingday	whether the day is neither a weekend nor holiday
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
temp	temperature in Celsius
atemp	feels like temperature in Celsius
humidity	relative humidity
windspeed	wind speed
casual	number of non-registered user rentals initiated
registered	number of registered user rentals initiated
Bike count	number of total rentals

Exploratory Data Analysis – Numeric Variables

Correlation Matrix

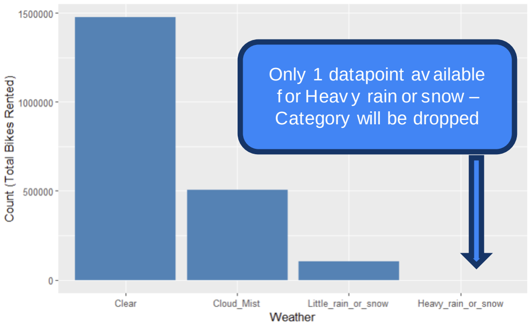


Key Takeaways

- **Temp and Atemp** have a high positive correlation among themselves and with the bike count.
- **Humidity** has a small negative correlation with bike count
- **Windspeed** has a very small correlation with the bike count
- **Casual and Registered** are highly correlated with bike_count because of the reason discussed earlier that they are components of the target variable (causal + registered = bike_count)

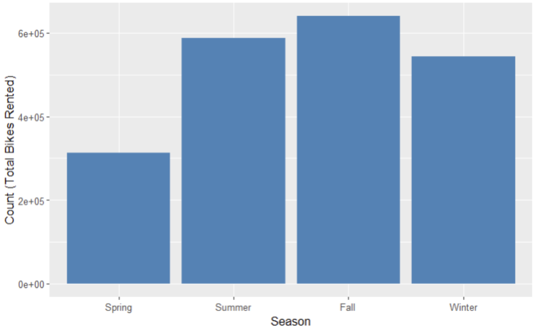
Exploratory Data Analysis – Categorical Variables (1/2)

Bike Count vs Weather



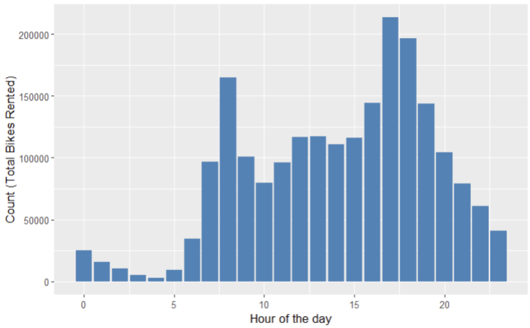
Clearer the weather, higher is the bike rental counts

Bike Count vs Season



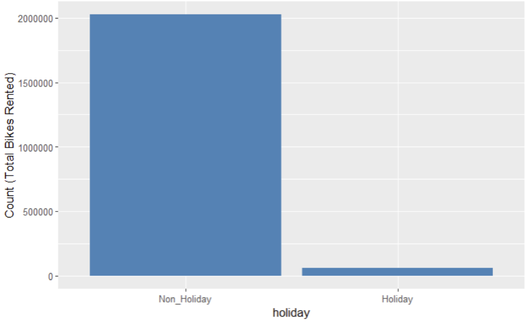
Higher counts during summer and fall, lower counts during winter and spring

Bike Count vs Hour of the Day



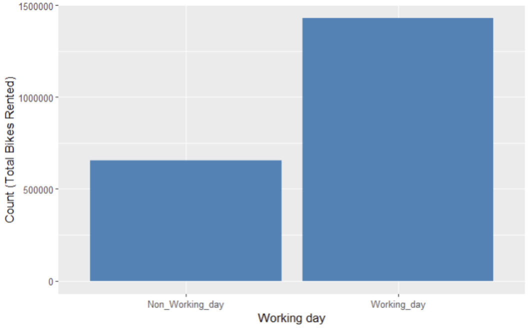
Peak counts during rush office hours, low counts during night, medium counts during rest of the day

Bike Count vs Holiday



Very low counts on holidays

Bike Count vs Working Day



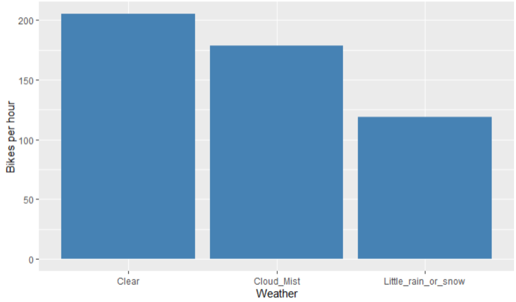
Higher counts on working days vs non-working days

Key Takeaway

Distribution of total # of bikes rented varies quite a bit by each of the categorical variables. This **needs to be validated by checking the average # of bikes rented per hour** vs each of the variables.

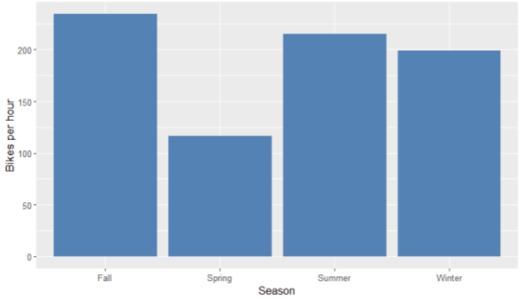
Exploratory Data Analysis – Categorical Variables (2/2)

Bikes per hour vs Weather



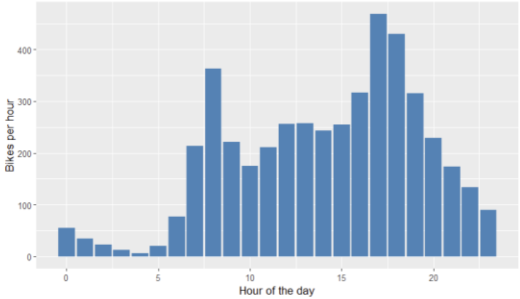
Clearer the weather, higher is the bike rentals per hour

Bikes per hour vs Season



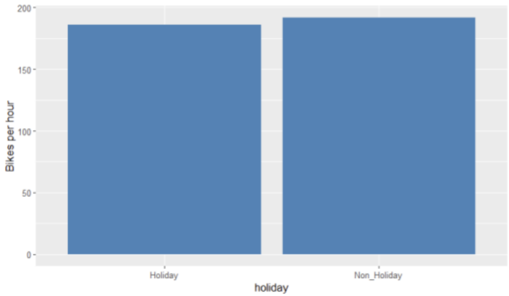
Higher bike rentals per hour during summer and fall, lower during winter and spring

Bikes per hour of the Day



Peak counts during rush office hours, low counts during night, medium counts during rest of the day

Bikes per hour vs Holiday



Holidays seem to have very little impact on bike rentals per hour

Bikes per hour vs Working Day

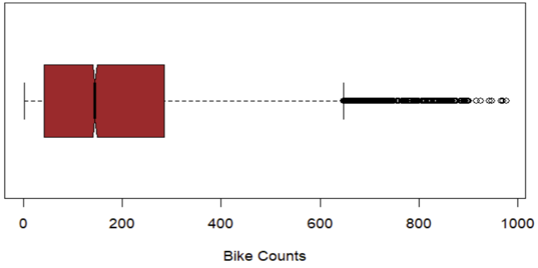
Working / Non-working days seem to have very little impact on bike rentals per hour

Key Takeaway

Except for holiday and working day, all other variables seem to have impact on the target variable, i.e., bikes rented per hour

Data Preparation

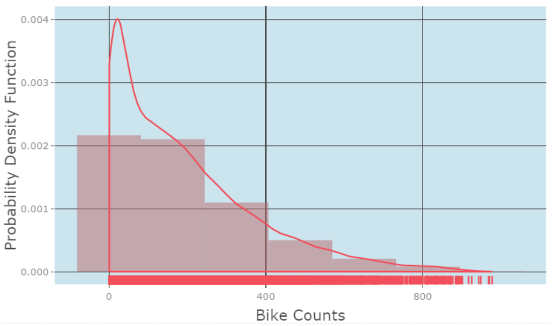
Identifying And Excluding Outliers



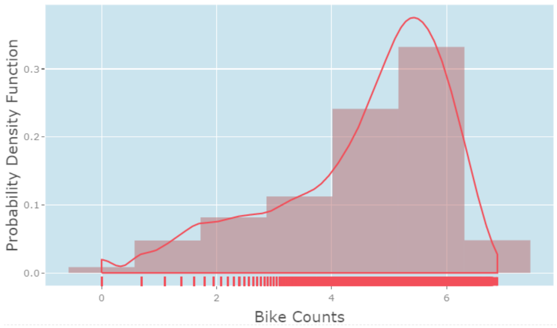
It is clear that there are a **few outliers** in the target variable. We will **remove records that fall outside of mean +/- 3 SD confidence interval** for the purpose of modelling.

Log Transformation of Target Variable

Original Distribution



After Log Tranformation



The distribution of target variable is **right-skewed.** To correct the right-skew we apply **logarithmic transformation**. The distribution looks much better after the transformation.

One Hot Encoding of Categorical Variables

In one hot encoding, we convert each categorical value of a variable into a **new categorical column and assign a binary value of 1 or 0** to those columns. *See example to the right:*

Season
Spring
Summer
Fall

One Hot Encoding →

Spring	Summer	Fall
1	0	0
0	1	0
0	0	1

Scaling Numeric Variables

Finally, we scaled the numeric variables as per:

(x – min(x))/ Range(x)

ANOVA Testing for Categorical Variables

```
Analysis of Variance Table

Response: log_bike_count
                      Df  Sum Sq Mean Sq   F value    Pr(>F)    
seasonSpring           1  1054.4 1054.43 2654.7093 < 2.2e-16 ***
seasonSummer           1     3.5    3.46    8.7013  0.003187 ** 
seasonFall             1    46.5   46.51  117.0952 < 2.2e-16 ***
weatherClear           1   120.8  120.78  304.0739 < 2.2e-16 ***
weatherCloud_Mist      1   178.6  178.56  449.5640 < 2.2e-16 ***
holidayHoliday         1     0.0    0.03    0.0859  0.769503    
workingdayWorking_day  1    17.0   16.99   42.7742 6.425e-11 ***
hour                  23 16827.1  731.61 1841.9558 < 2.2e-16 ***
day                   18    13.8    0.76    1.9255  0.010523 *  
month                  8   205.5   25.69   64.6683 < 2.2e-16 ***
year                   1   734.7  734.69 1849.7092 < 2.2e-16 ***
Residuals          10680  4242.0    0.40                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Key Takeaways

- Variables on **season, weather, workingday, hour, day, month and year** have very low p-values indicating that they are all **significantly associated** with the target variable log_bike_count
- On the other hand, variable - **holiday** has slightly higher p-value, suggesting that they **may not be statistically significant** in explaining the variance of log_bike_count
- **For Month, the degree of freedom should have been 11 instead of 8 (Due to singularities).** Hence, we decided to drop it, to avoid confusion

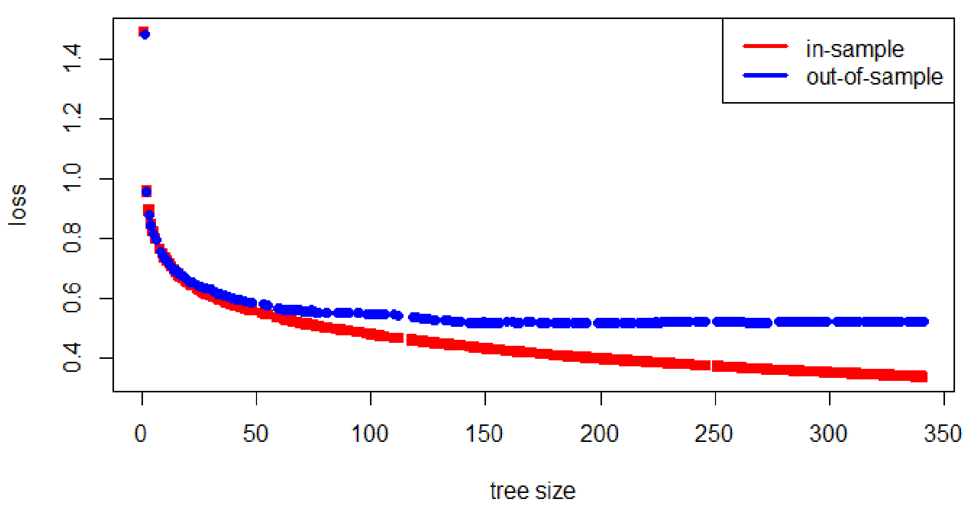
Modeling Results – Linear Regression

Step No.	Step Description	Key Metrics	Key Takeaways
1	**Base Linear Model** - Throw in all variables to create a base linear model	• **Residual standard error:** 0.626 • **Multiple R-squared:** 0.821 • **Adjusted R-squared:** 0.820	• **Days** variables show up as **not significant** in explaining the target variable • **VIF** indicates **strong collinearity** between **temp and atemp – Dropping atemp**
2	**Lasso and Ridge** Regression – to regulate the existing variables	• **Lambda_Lasso:** 0.00096 • **Lambda_Ridge:** 0.054	• **For both Lasso and Ridge**, we see similar results: while Lasso was not able to reduce coefficient of any variable to exactly 0, Ridge too **did not severely punish coefficient of any variable**
3	**2nd linear model** – to predict without using the insignificant variables identified in the above two steps	• **Residual standard error:** 0.627 • **Multiple R-squared:** 0.821 • **Adjusted R-squared:** 0.820	• While residual std. error deteriorated marginally, multiple R^2 and **adjusted R^2 remained unchanged** • However, there is **not much difference from the base model**
4	**Forward and Backward Regression** – to identify the best set of predictors and to check if it helps in better estimation	• **AIC:** -9941.06 • **Multiple R-squared:** 0.821 • **Adjusted R-squared:** 0.820	• **The best AIC is achieved** when **ALL** the variables (other than those excluded above) are included • **Multiple R^2 and adjusted R^2 are the same** as achieved in the previous step

Overall, linear model does not seem to be the ideal solution for this problem as the model barely improves through various iterations

Modeling Results – Regression Trees

Tree size vs Loss (In-sample vs Out of Sample)

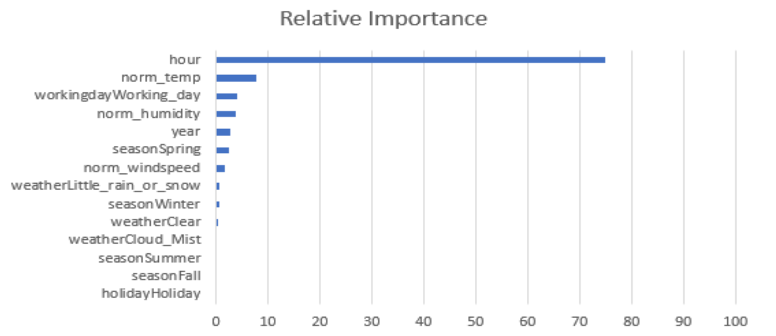


Metric	Value
Big tree size	341
Optimal tree size (After pruning)	139
CP value (Optimal)	0.0002
Loss (out of sample)	**0.515**

Modeling Results – Random forest and Boosting

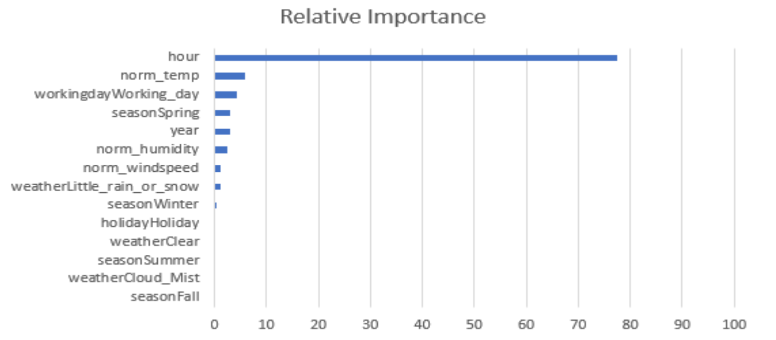
Random Forest

mtry <dbl>	ntree <dbl>	olrf <dbl>	ilrf <dbl>
5	500	0.427	0.435
10	500	0.402	0.402
5	1000	0.426	0.434
10	1000	0.403	0.402



Boosting

tdepth <dbl>	ntree <dbl>	lam <dbl>	olb <dbl>	ilb <dbl>
8	1000	0.01	0.370	0.342
10	1000	0.01	0.363	0.329
8	5000	0.01	0.363	0.261
10	5000	0.01	0.365	0.244
8	1000	0.02	0.359	0.307
10	1000	0.02	0.358	0.295
8	5000	0.02	0.373	0.221
10	5000	0.02	0.373	0.200



The best out of sample mean square error (0.358) is achieved through boosting when the parameters are:

- Depth = 10
- # trees = 1000
- Shrinkage factor = 0.02

Recommendations and Conclusion

Key Takeaways

- Bikes should be highly stocked at **8-9AM** and **4-7PM**, which are peak commuting hours, and the times we expect the most bike rentals.
- People are more likely to rent bikes during **moderate temperatures** and **clear weather**. Service providers should expect higher demand during these days, and **lower during the spring season**.
- **Working day status is a strong predictor** of bike rentals, even though in EDA, we didn't find it to be of significance as a stand-alone predictor. This is because of the interactions that working day status has with other variables.
- Findings are limited to DC, as other cities may not be as walkable or have the same seasons