Analyzing NYC Citi Bike System

Will It Be A Busy Day For Citi Bikes Today ?

Miners “R” Us

AGENDA

- The Data
- Data Preprocessing
- Exploratory Data Analysis
  - Descriptive Analysis & Summary
  - Data Visualization
- Predictive Analysis
  - Decision Tree
  - Logistics Regression Model

INTRODUCTION

- Citi Bike is the most widely used **bike sharing program** in New York City and Jersey City.

- Citi Bike was proposed in an effort to
  - **reduce emissions**
  - **reduce collisions and road transit congestion**
  - **improve public health**

Scope

- New York City Boroughs
- Citi Bike Stations and Routes

Goal

- To predict the usage of Citi Bikes based on demographic variables available in the CitiBike dataset.

The Data

- 2016 Citi Bike Dataset - https://www.citibikenyc.com/system-data
  - Observation – Sequential order of each trip taken
  - Variables – Trip Duration, Start Time and Date, Stop Time and Date, Start Station Name, End Station Name, Station ID, Bike ID, User Type, Gender, Year of Birth
- 2016 Weather Dataset – Kaggle
  - Observation - Sequential order of weather parameters for each day
  - Variables – maximum temperature, minimum temperature, average temperature, precipitation, snow fall, snow depth

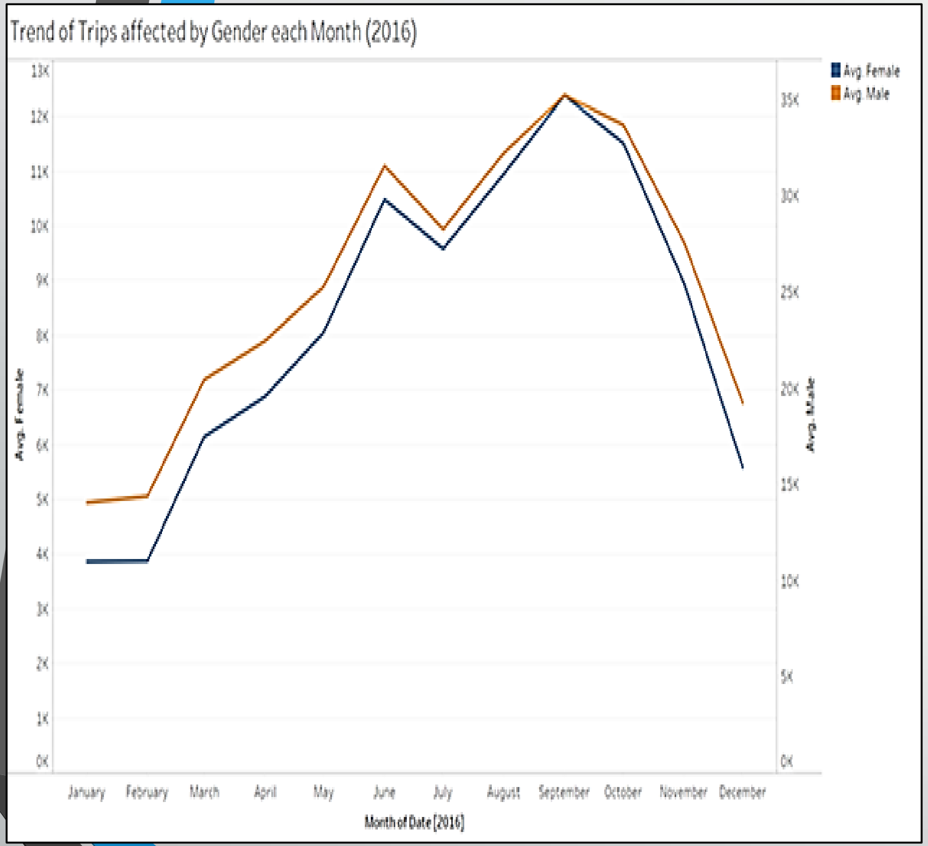
Data Pre Processing

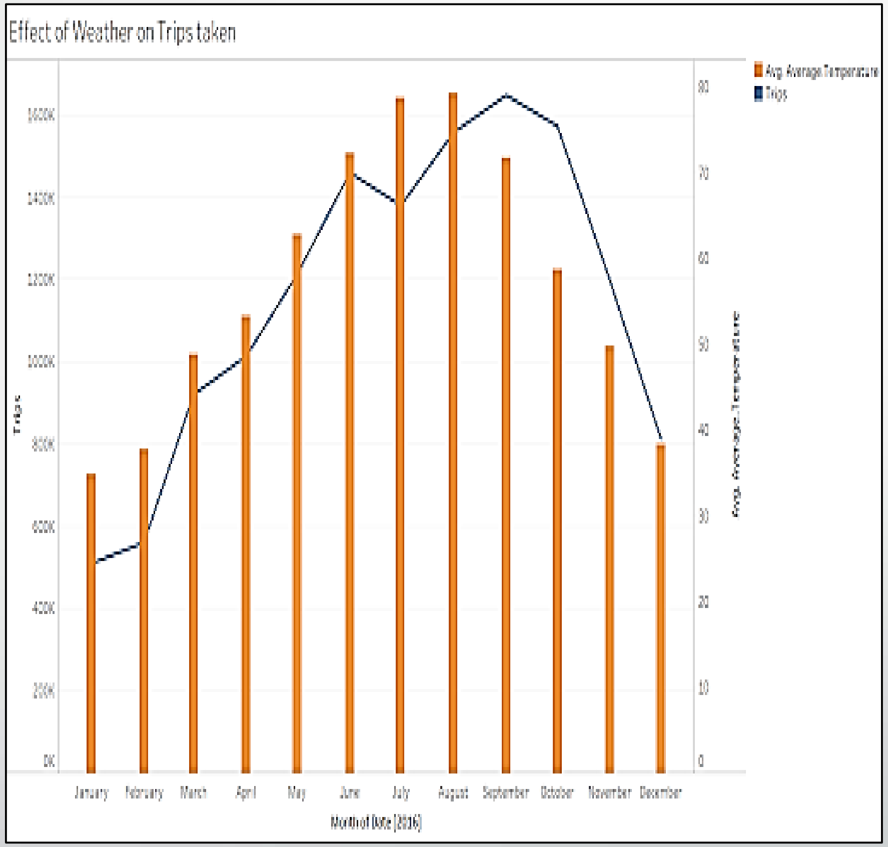
- Replaced Null Values with Mean
- Aggregation Function – Gender, User Type, Age
- Aggregation of Weather Dataset
- Indicators – Weekend & TripsMorethan38K

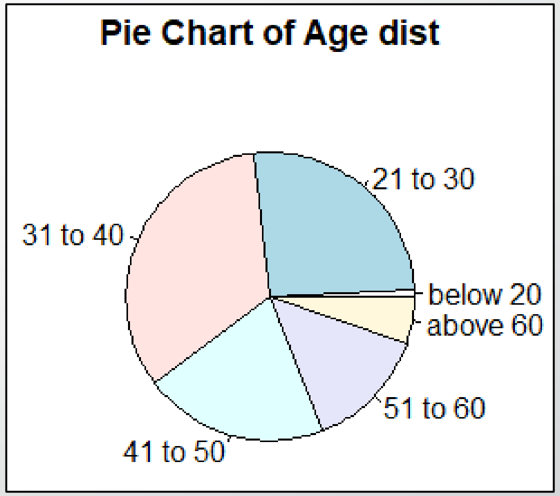
Aggregated Dataset

Numerical	Categorical
Date	WeekendIndicator
Male	Days
Female	Morethan38K
Subscriber	
Customer	
Maximum temperature	
Minimum temperature	
Average temperature	
Snow fall	
Age Groups	
Trips	
Miles Travelled	

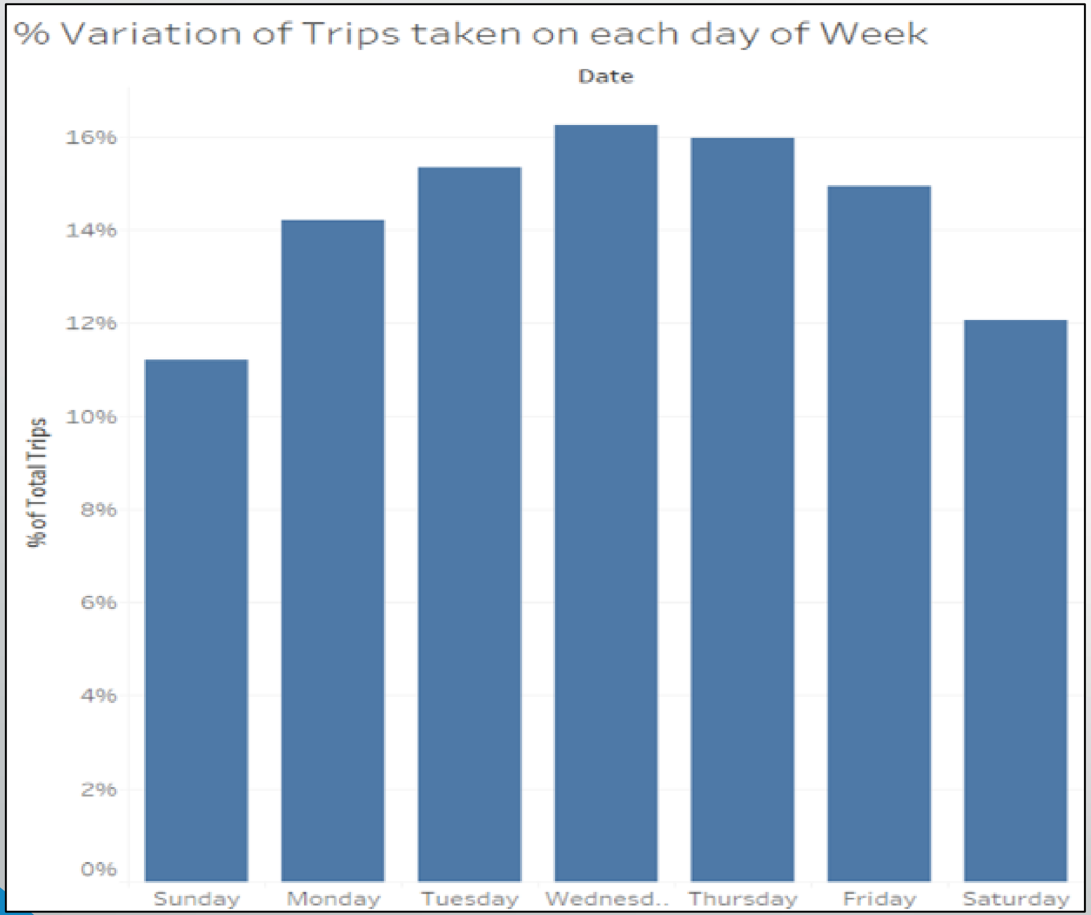
Exploratory Analysis

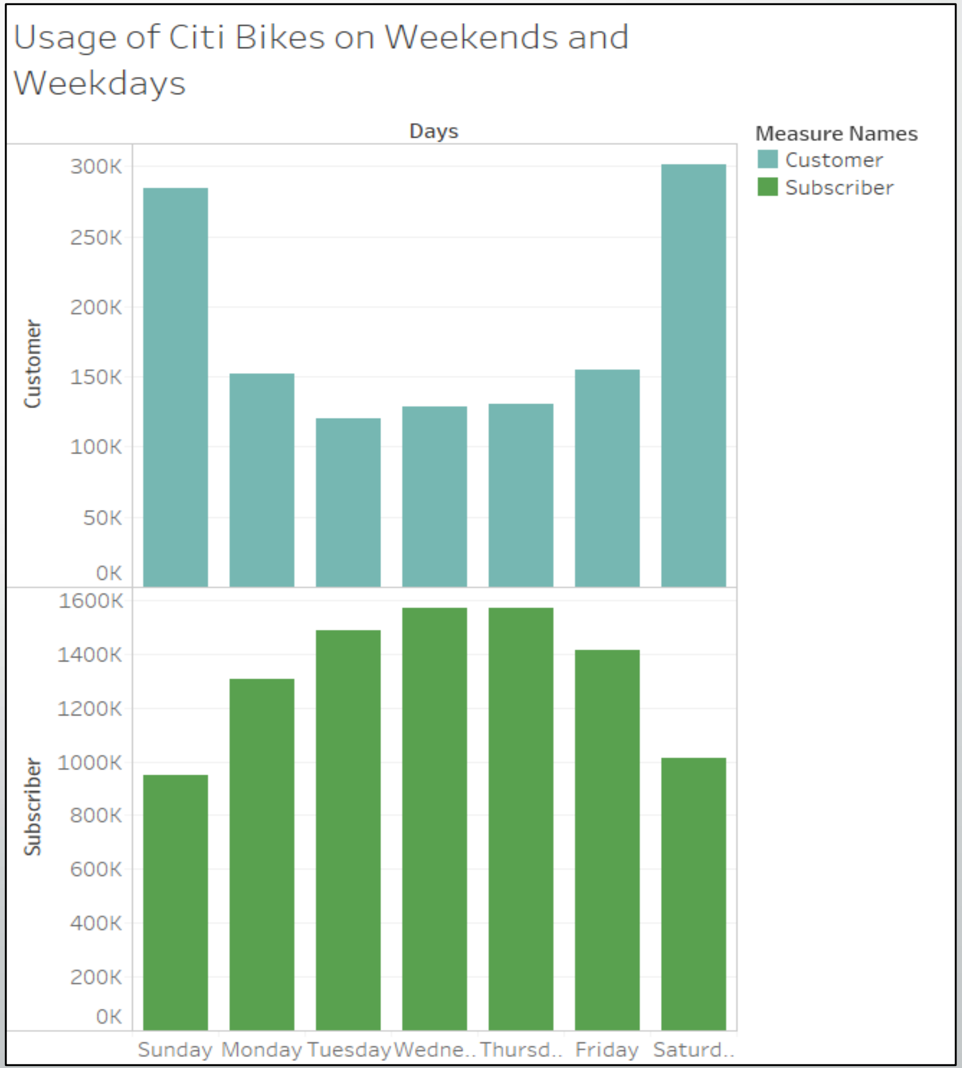






Exploratory Analysis





Predictive Analysis – Logistic Regression Model

Predict whether usage of Citi Bikes will me more than 38K

```
Call:
glm(formula = Morethan38K ~ WeekendIndicator + average.temperature +
    Subscriber + Customer, family = binomial(link = "logit"),
    data = trainingData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.86730  -0.55718   0.06765   0.57432   1.94940

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -7.583e+00  1.027e+00  -7.385 1.52e-13 ***
WeekendIndicator1   -8.260e-01  4.189e-01  -1.972   0.0486 *
average.temperature  1.474e-01  2.105e-02   7.001 2.54e-12 ***
Subscriber          -5.409e-06  1.432e-05  -0.378   0.7057
Customer            -1.048e-04  6.752e-05  -1.553   0.1205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.89  on 255  degrees of freedom
Residual deviance: 205.78  on 251  degrees of freedom
AIC: 215.78

Number of Fisher Scoring iterations: 5
```

- **AIC Value (Akaike Information Criteria)**
  - WeekendIndicator and AverageTemperature the AIC value was 220.39
  - WeekendIndicator, AverageTemperature, Subscriber and Customer the AIC value lowered to 215.78
- **Residual Deviance Value**
  - WeekendIndicator and AverageTemperature the Residual value was 209.26.
  - WeekendIndicator, AverageTemperature, Subscriber and Customer the Residual value lowered to 205.78.

Confusion Matrix

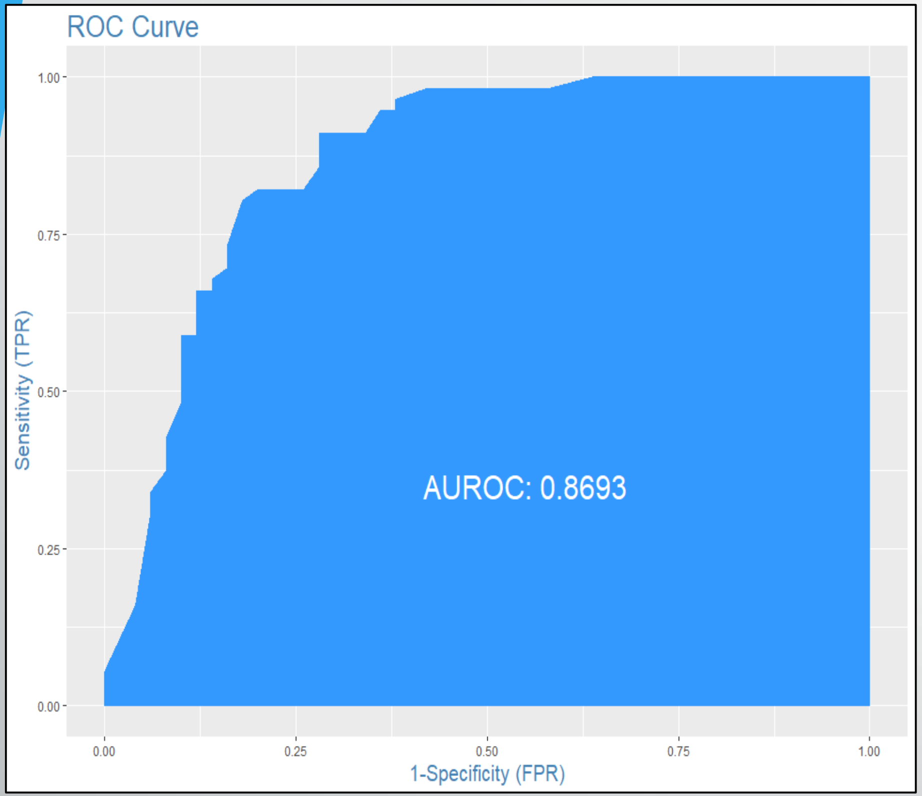
		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

```
> confusionMatrix
   0  1
0 36  5
1 14 51
```

Accuracy of Model = (TN + TP) / (TN+TP+FN+FP)

Accuracy = 0.8207547

ROC (Receiver Operating Characteristic) Curve



Area under ROC curve is 86.93%
which is pretty good

```
accuracy(testData$Morethan38K,predicted,threshold= optCutOff)
threshold       AUC omission.rate sensitivity specificity prop.correct     Kappa
0.3833166 0.8153571    0.08928571   0.9107143        0.72    0.8207547 0.6368554
```

Truth detection rate of 91% on test data is good

Predictive Analysis – Decision Tree

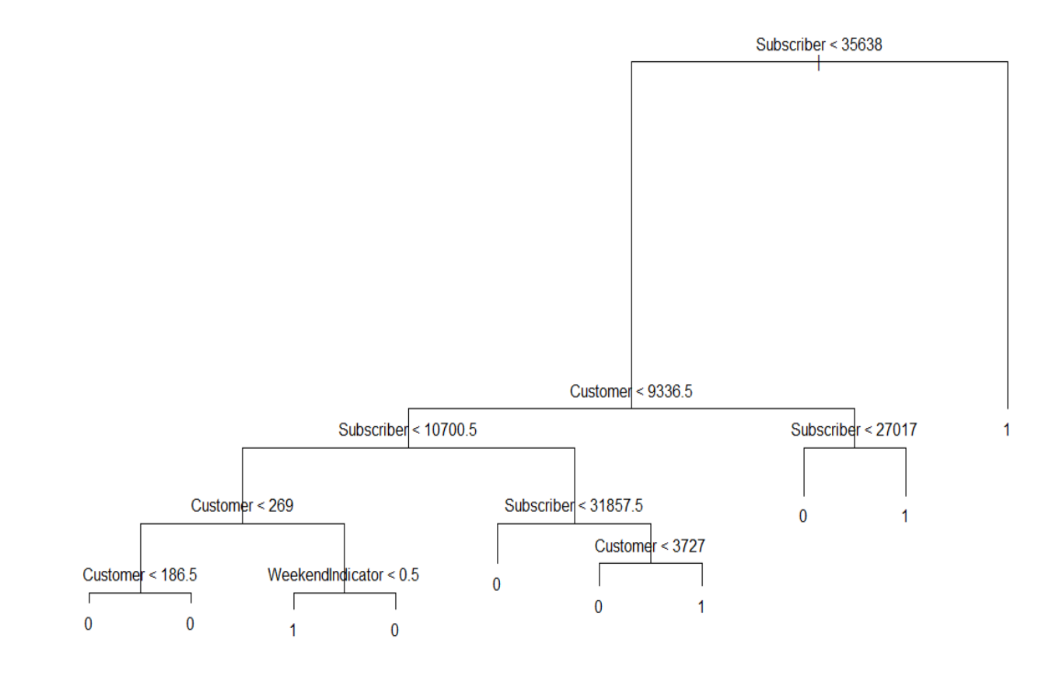
- Using 2 Decision Tree models, we were able to select which variables best predict usage (MoreThan38K).

Model 1:

- Input Variables: Customer + Subscriber + WeekendIndicator (no Average Temperature)
- Label: MoreThan38K

```
Classification tree:
tree(formula = train_labels ~ Customer + Subscriber + WeekendIndicator,
    data = train_data)
Number of terminal nodes:  10
Residual mean deviance:  0.2387 = 68.04 / 285
Misclassification error rate: 0.06102 = 18 / 295
```

```
              test_data.test_labels
my.predictions  0  1
             0 26  9
             1  0 32
Confusion Matrix and Statistics

              test_data.test_labels
my.predictions  0  1
             0 26  9
             1  0 32

               Accuracy : 0.8657
                 95% CI : (0.7603, 0.9367)
    No Information Rate : 0.6119
    P-Value [Acc > NIR] : 4.729e-06

                  Kappa : 0.734
 Mcnemar's Test P-Value : 0.007661

            Sensitivity : 1.0000
            Specificity : 0.7805
         Pos Pred Value : 0.7429
         Neg Pred Value : 1.0000
             Prevalence : 0.3881
         Detection Rate : 0.3881
   Detection Prevalence : 0.5224
      Balanced Accuracy : 0.8902

       'Positive' Class : 0
```



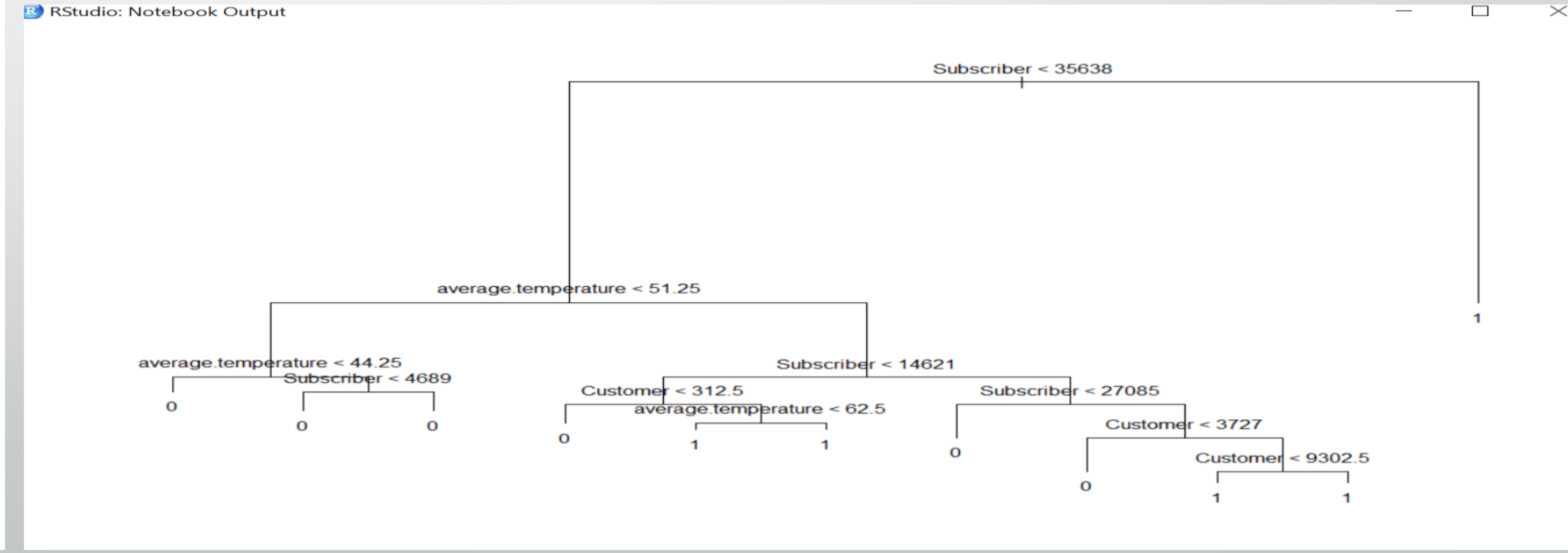
Predictive Analysis – Decision Tree

Model 2: How will accuracy change with the following variables?

- Input variables: Customer + Subscriber + Average Temperature + WeekendIndicator
- Label: Still MoreThan38K

```
Classification tree:
tree(formula = train_labels ~ Customer + Subscriber + average.temperature +
    WeekendIndicator, data = train_data)
Variables actually used in tree construction:
[1] "Subscriber"          "average.temperature" "Customer"
Number of terminal nodes:  11
Residual mean deviance:  0.145 = 41.17 / 284
Misclassification error rate: 0.04068 = 12 / 295
```

```
my.predictions  0  1
             0 24  6
             1  2 35
Confusion Matrix and Statistics

              test_data.test_labels
my.predictions  0  1
             0 24  6
             1  2 35

               Accuracy : 0.8806
                 95% CI : (0.7782, 0.947)
    No Information Rate : 0.6119
    P-Value [Acc > NIR] : 1.097e-06

                  Kappa : 0.7555
 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.9231
            Specificity : 0.8537
         Pos Pred Value : 0.8000
         Neg Pred Value : 0.9459
             Prevalence : 0.3881
         Detection Rate : 0.3582
   Detection Prevalence : 0.4478
      Balanced Accuracy : 0.8884

       'Positive' Class : 0
```



Conclusion

- The prediction of Usage is based on Customer, Subscriber, WeekendIndicator and AverageTempearture.
- Using Logistic Regression, the accuracy rate is 82%
- Using the decision tree confusion matrix, the accuracy of our model is 88%.
- Using EDA and Predictive Models, we find that user type (Subscriber & Customer), Day of Week(Weekend Indicator), are the best indicators of usage for a given day.
- Citi Bike can optimize their maintenance resources to improve stations and bikes during anticipated higher-usage days.

Thank You!