Seoul Bike Dataset

IE 590 ML Group 6:
Andra Yahya, Jayendran Ravindran, Olivia Wang, Roberto Herrera

PURDUE UNIVERSITY | School of Industrial Engineering

Problem Statement

Overview, Questions, and Justification

Overview	Questions	Justification
• Rental bikes provide accessibility and urban mobility to Seoul citizens • Need to ensure enough rental bikes available at the right time for the public. • Predict the rental bike count required at each hour using ML techniques.	• To predict the rental bike count required at each hour using attributes. • To determine the attributes that play a major role in predicting the rental bike count. • To provide rental bikes to all citizens who are in dire need of bikes are covered.	• Enough rental bikes supplied so that residents are not inconvenienced. • Use the limited resources of Seoul corporation frugally by not supplying excess bikes. • Maintenance and repair of bikes can be planned by knowing the right rental bike count.

Problem Statement

Benefits

Seoul Metropolitan Corporation	Seoul Citizens	Rental Bike Service Providers
• Can allocate the right amount of money and resources for supplying rental bikes and cut waste by not supplying excess bikes.	• Can travel safely and conveniently within the city of Seoul because of the availability of rental bikes at each hour. • Can provide accessibility and mobility to citizens without private transportation.	• Can plan their daily supply of rental bikes from inventory based on demand. • Can also plan their routine maintenance and repair of bikes.

Problem Statement

Specific Goals and Subgoals

Data collection	• Data collection & cleaning • Feature selection • Identify/handle missing/inconsistent data	Clean and consistent data
Model Development	• Training & test data split • Regression model development using training dataset	High R-squared, Low MSE
Model Evaluation	• Regression model evaluation using test dataset • Minimizing test error using K-fold cross validation	High accuracy, Low variance
Model Improvement	• Ridge and Lasso regression for avoiding potential overfitting • Optimal tuning parameters and best model selection using CV • PCA for selecting important components	Higher R-squared (>0.65), Lower MSE

Methods

Preliminary Methods

- Primary Method: Linear Regression
  - Models the relationship between hourly rented bike count (response variable) and weather and date (predictor variables)
- Model Evaluation Metrics
  - Mean Squared Error (MSE)
  - R-squared
- Additional Techniques
  - K-Fold Cross Validation
  - Ridge Regression
  - Lasso Regression

Methods

Revised Methods

- Wrong Nonlinear Assumption
  - Found high RMSE values that indicated poor fit
  - Visual analysis, through correlation & scatter plots, reveals a nonlinear relationship between features and bike rental count
- Transition to Non-Linear Modeling to enhance prediction accuracy
  - Polynomial Regression
  - Piecewise Constant Regression
  - Splines

Results

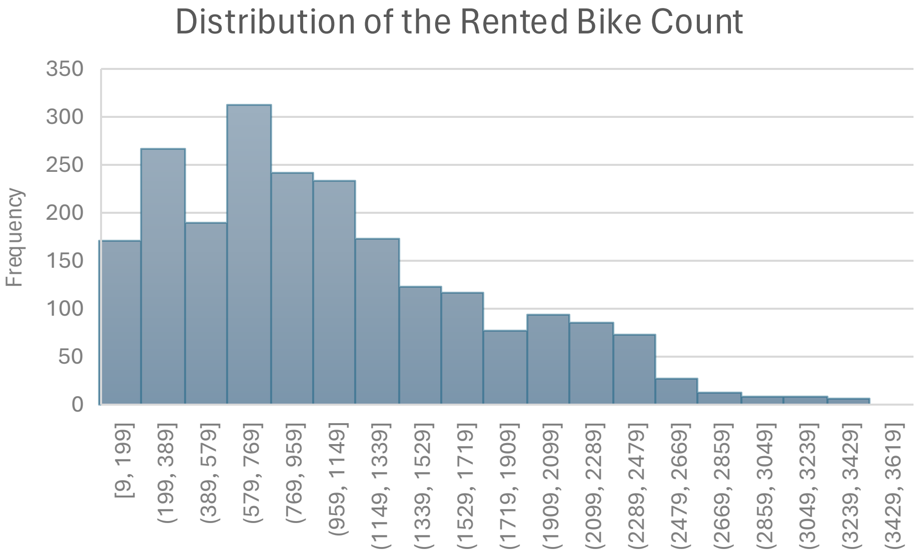
Linear models

```
> summary(lm_model)

Call:
lm(formula = Rented.Bike.Count ~ . - Date, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1145.9  -278.2   -56.0   213.8  2246.1 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               2.565e+04  3.543e+03   7.239 5.01e-13 ***
Hour                      2.723e+01  8.267e-01  32.946  < 2e-16 ***
Temperature..C.           2.044e+01  4.036e+00   5.066 4.17e-07 ***
Humidity...              -1.001e+01  1.122e+00  -8.918  < 2e-16 ***
Wind.speed..m.s.          1.928e+01  5.705e+00   3.380 0.000728 ***
Visibility..10m.          1.226e-02  1.112e-02   1.103 0.270264    
Dew.point.temperature..C. 7.326e+00  4.211e+00   1.740 0.081946 .  
Solar.Radiation..MJ.m2.  -8.448e+01  8.555e+00  -9.874  < 2e-16 ***
Rainfall.mm.             -6.000e+01  4.686e+00 -12.805  < 2e-16 ***
Snowfall..cm.             3.221e+01  1.270e+01   2.535 0.011260 *  
SeasonsSpring            -4.065e+02  3.986e+01 -10.200  < 2e-16 ***
SeasonsSummer            -3.006e+02  2.700e+01 -11.136  < 2e-16 ***
SeasonsWinter            -7.626e+02  5.851e+01 -13.034  < 2e-16 ***
HolidayNo Holiday         1.286e+02  2.436e+01   5.278 1.35e-07 ***
Functioning.DayYes        9.433e+02  3.027e+01  31.157  < 2e-16 ***
DateNumeric              -1.449e+00  1.987e-01  -7.295 3.31e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 434.5 on 6993 degrees of freedom
Multiple R-squared:  0.551,	Adjusted R-squared:   0.55 
F-statistic: 572.1 on 15 and 6993 DF,  p-value: < 2.2e-16
```

Summary of Linear Regression Model

Ridge regression and Lasso regression:

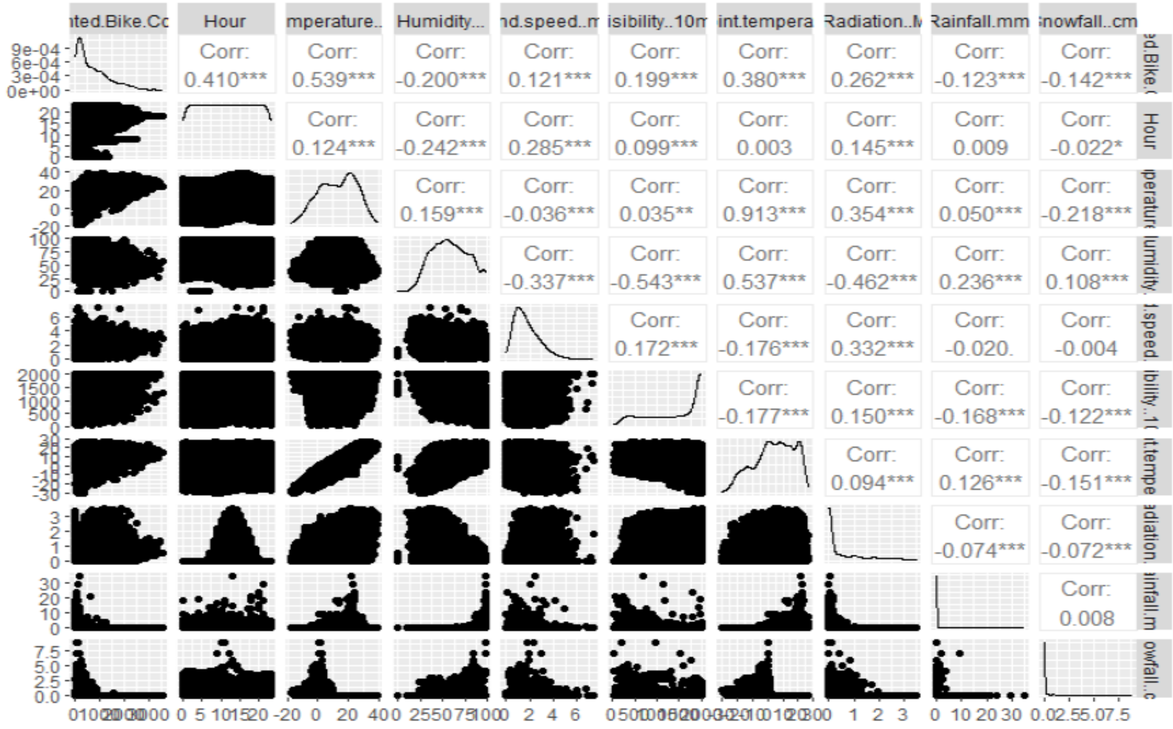
	Ridge regression	Lasso regression
RMSE	433.39	432.46
R-Squared	0.5484	0.5503



Standard Deviation: 690.24
Variance: 476,437.83
Range: 3,547

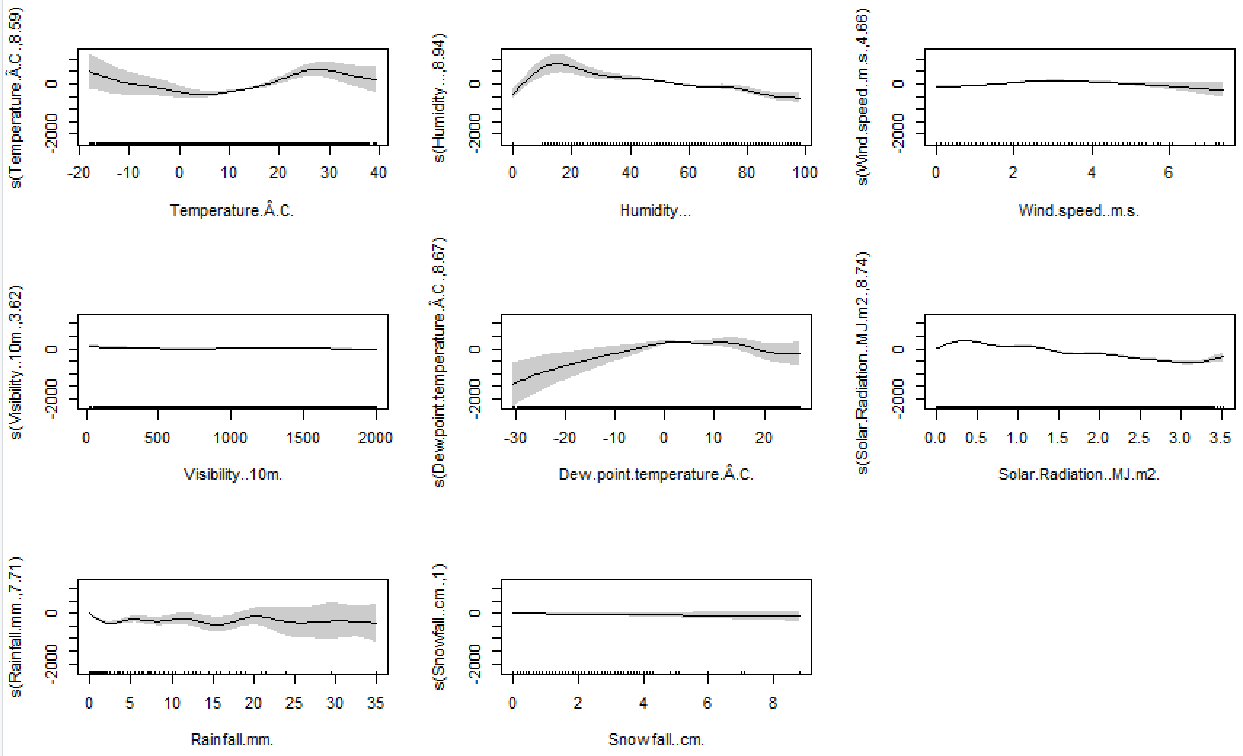
Results

Correlations between variables:



GGpairs correlation plots for numeric features

GAM

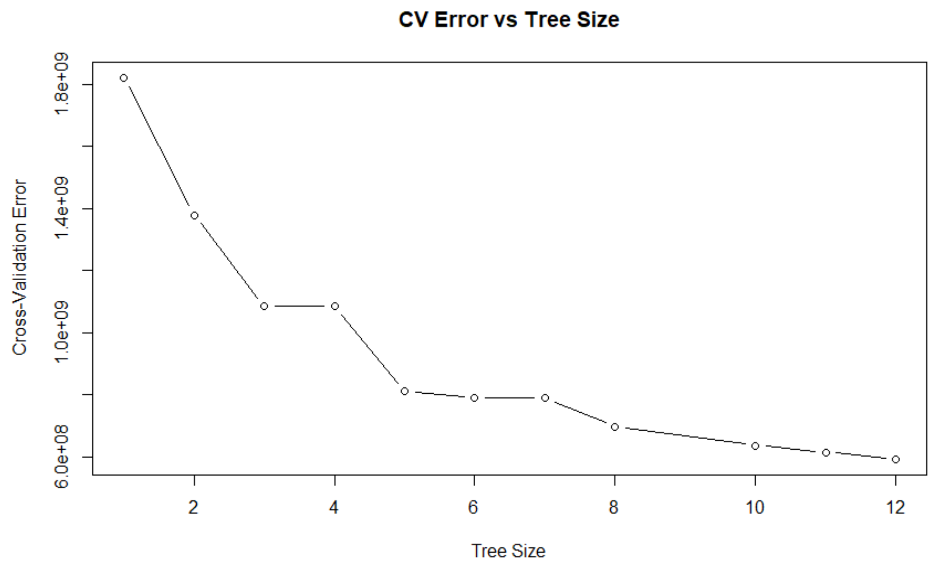


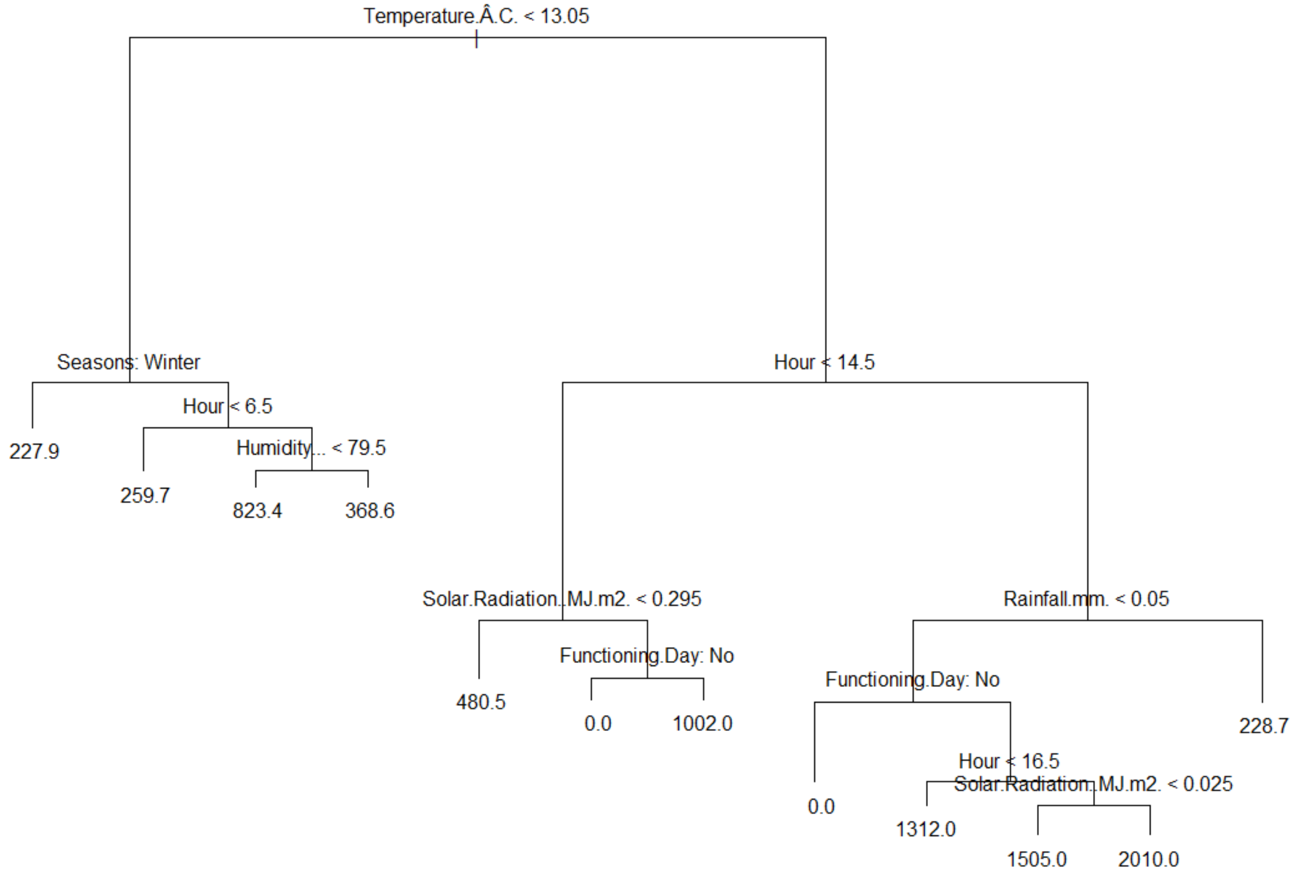
Need to try non-linear models

Results

Non-linear models

	GAM	Decision Tree	Random forest
RMSE	404.61	354.44	222.11



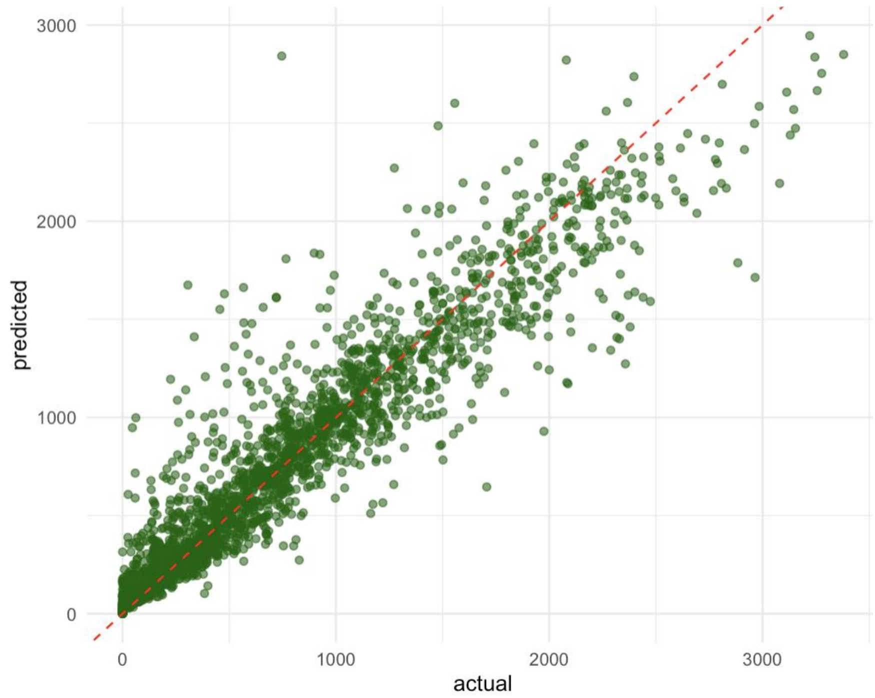


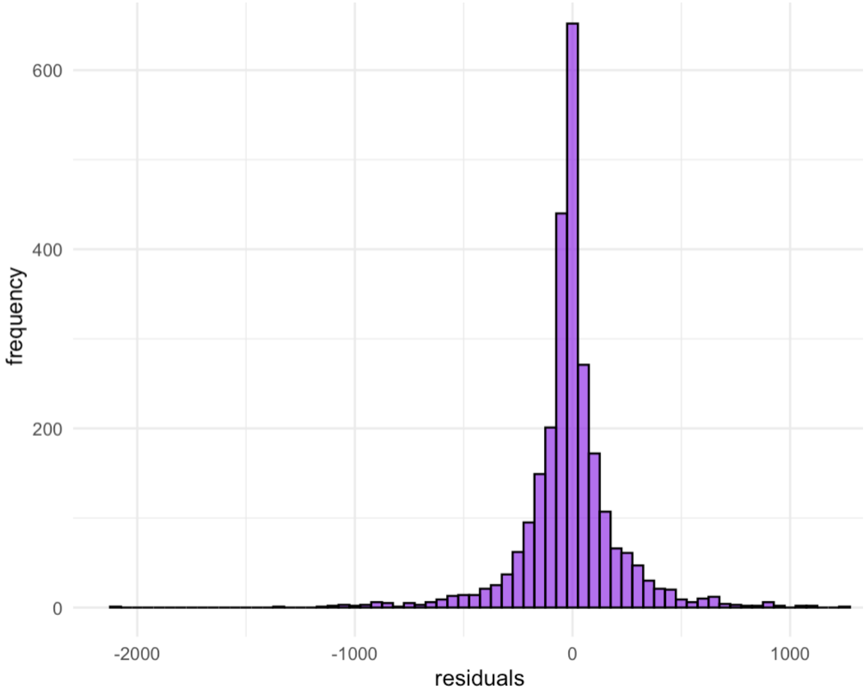
Results

Non-linear models

Random forest

```
> importance(rf_model)
                          %IncMSE IncNodePurity
Hour                    147.76230     874664532
Temperature              47.38723     499346814
Humidity                 56.79201     251506742
Wind.speed               23.43699      66371703
Visibility               27.41686      72734473
Dew.point.temperature    26.94678     157030088
Solar.Radiation          37.56240     169745786
Rainfall                 46.12981     111490430
Snowfall                 13.58927       2871193
Seasons                  20.96415     149283635
Holiday                  24.20762       8158572
Functioning.Day         123.97982     206081881
Month                    36.26367     318417278
```





Lessons Learned

Summary

- We used **Linear Method** but got high MSE results.
  - Linear Regression, K-fold cross validation, Ridge, and Lasso
- We check for linearity assumption on the data set and found out most features are **non-linear**.
- We decide to go with **non-linear model**: Random Forest have the lowest RMSE = 222.11
  - Module 7: Moving Beyond Linearity
  - Module 8: Tree based Methods
- Why predicting bike rental counts can be **challenging?**
  - High Variability in Demand
  - Complex Seasonality
  - Randomness in Human Behavior
  - Nonlinear Interactions Between Variables
  - Seasonal Extremes and Outliers

Lessons Learned

Challenges & Solution

- **Challenges:**
  - The date column (day/month/year) feature is not R and Python friendly.
  - High RMSE value around 436 with linear model
  - Model exhibits non-linearity and underfitting
  - Categorical variables in the data set = seasons, holiday (yes/no), functioning day (yes/no)
- **How we will overcome the challenges:**
  - Convert the date column to **just months** to know if certain months also affect the rental bike count.
  - Given the categorical data, we will use **as.factor** function for the categorical
  - Given the non-linear and complex relationships observed and underfitting, we will be considering **non- linear models** that can capture these non-linearities
    - GAM (General additive model)
    - Decision tree
    - Random forest
    - Bagging
    - Boosting
  - We will do **different models for different seasons** since the behavior of the output variable quite different per seasons (Final report)
    - Seasonal: Summer, Winter, Spring, Autumn

PURDUE UNIVERSITY | School of Industrial Engineering

Thank You