Biking - An Emerging Alternative Transportation

Capstone project 2

Background Information

- A surge in bike riding in big cities for commuting to work, running errands and many more.
- Expanding rental systems - borrowing bikes from one station and return them to another station.
- Casual riders - the number of casual riders is increasing

Motivation

- More insights into this emerging phenomenon
- Potential stakeholders
  - Business firms: expanding and running business more efficiently
  - Local governments: riders’ security, biking lanes, and pollution free cities
  - Health professionals: avail the opportunity to communicate biking an avenue for better health

Approach

- Data Wrangling
- Discovering stories
- Statistical Analysis
- Algorithm Running
- Outcome Analysis
- Recommendation

Dataset Information

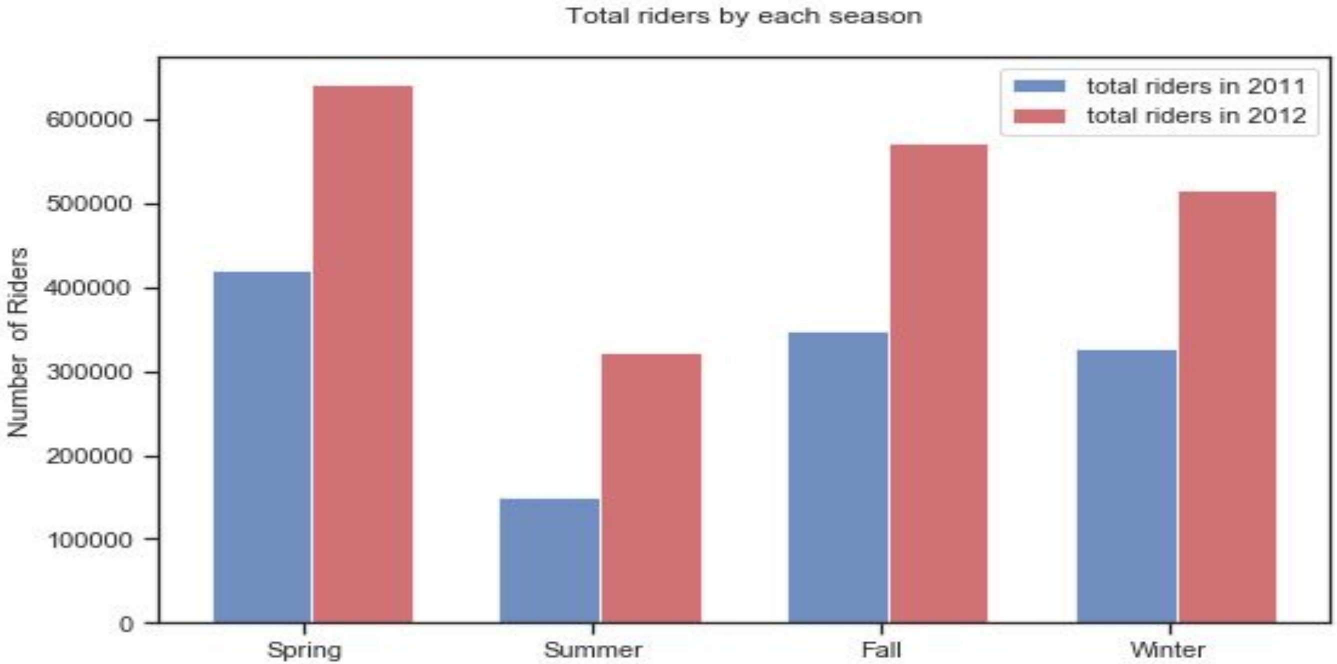
- The dataset used in this project is available at [Kaggle]
- The dataset provides relevant information for bike riders in Washington, D.C.
- This is a labelled dataset and The dataset has 13,379 rows and 18 columns
- The target variable is total number riders and will perform regression analysis
- The source code is available here.

Data Wrangling

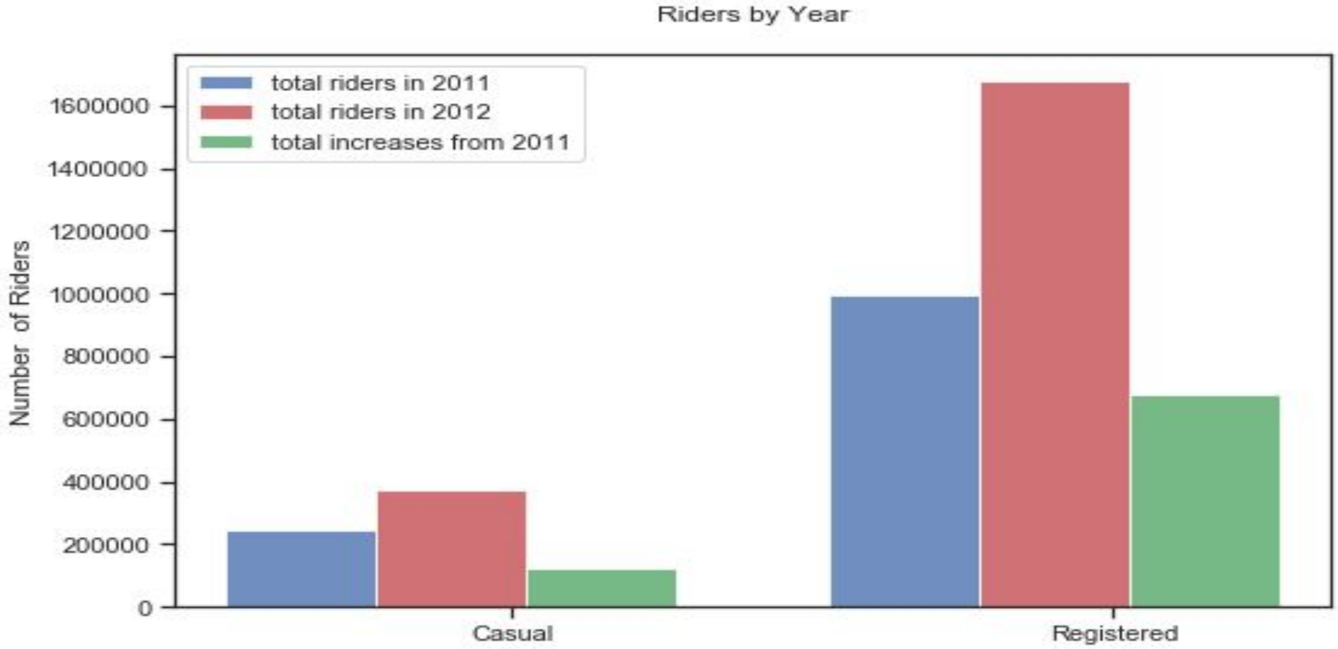
The data was clean:

- No missing values
- No outliers

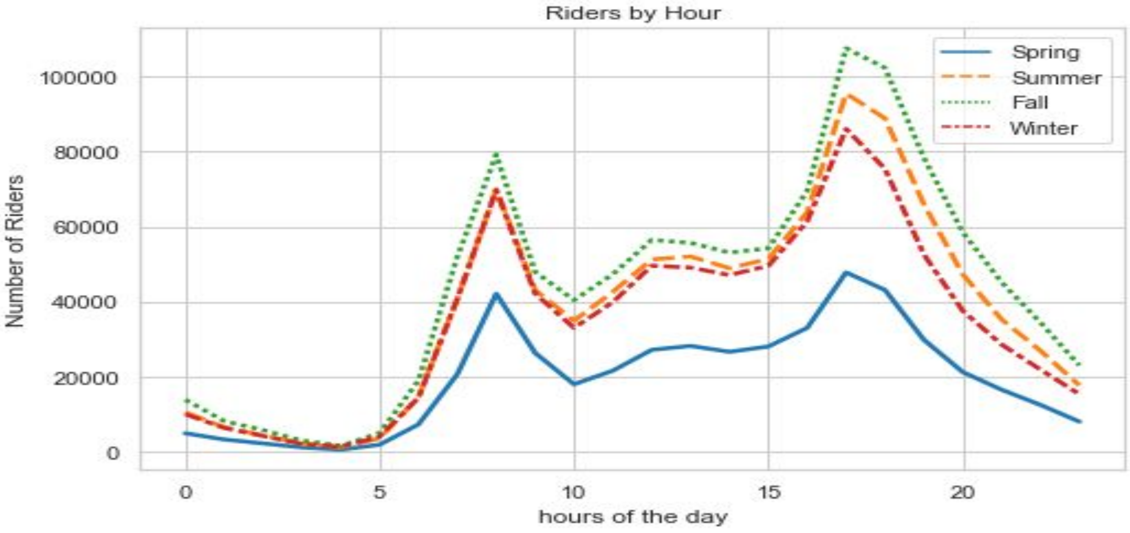
Data Story 1: The number of bike riders is on the rise.



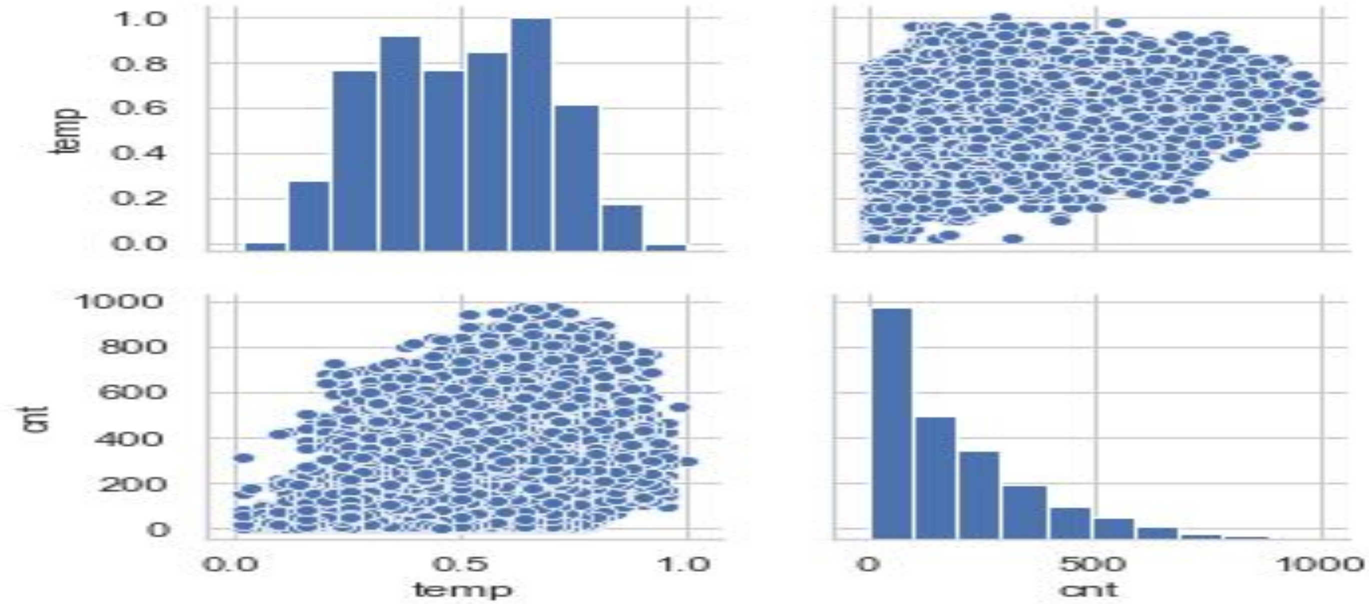
Data Story 2: The casual riders are increasing at a higher rate - 50% to 40%.



Data Story 3 : More than one hundred thousand bike riders on certain time of the day - rush hour around 8am and around 6pm.



Data Story 4: The temperature has no control on the number of bike riders



Statistical Analysis I

There is no difference between the number of riders in 2011 and the number of riders in 2012;

Null Hypothesis: There is no difference between the number of riders in 2011 and in 2012

Alternative Hypothesis: There is a difference between total riders in 2011 and in 2012

t = -22.160400484938148

p = 2.5301602945983702e-107

Statistical Analysis II

There is no difference between the number of casual and registered riders.

Null Hypothesis: There is no difference between casual riders and registered riders

Alternative Hypothesis: There is a difference between casual riders and registered riders

$t = -97.81332643791566$

$p = 0.0$

Machine learning

Data Readiness

- Taking a 70% threshold: dataset contains 11,643 training and 5,736 test instances
- Total number of attributes : 57
- Target variable: number of bike riders per hour

Algorithms

- Initial algorithms: linear regression, polynomial, ridge, lasso, kneighbors, support vector machine(SVM), decision trees, adaBoost, gradient boosting, neural network, and random forest.
- Final algorithms: linear regression, polynomial regression, ridge with polynomial data, random forest and support vector machine

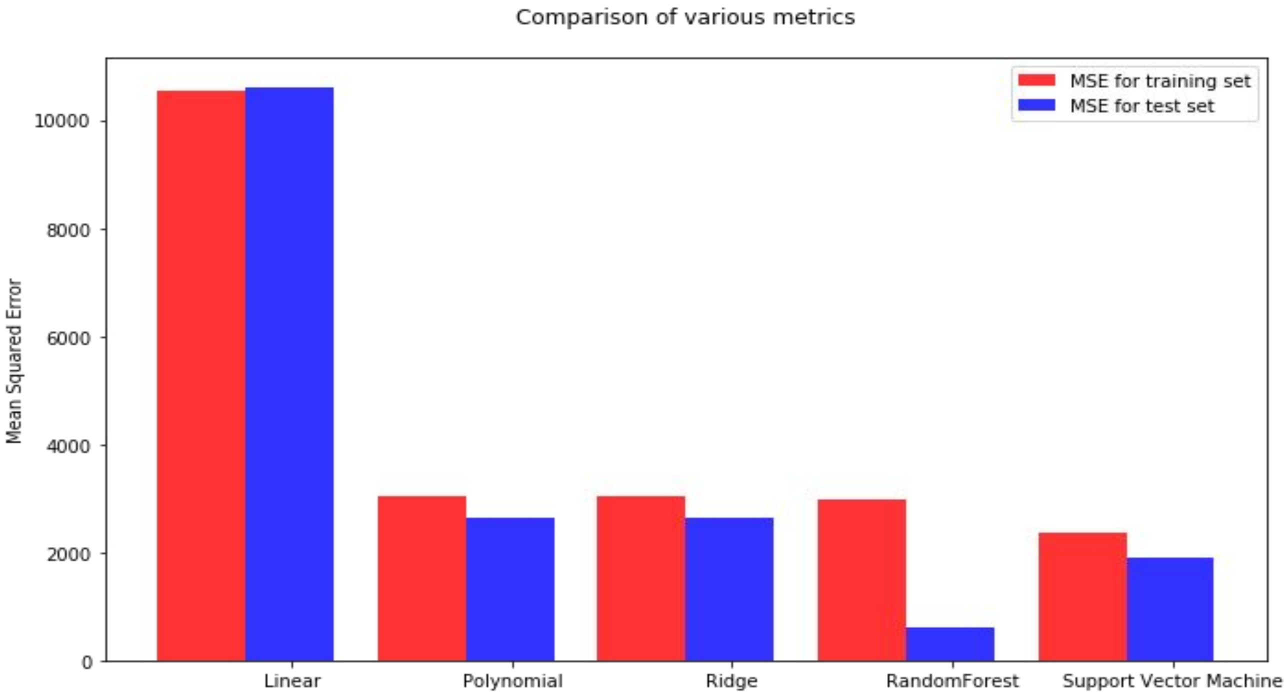
Evaluations

- Mean Squared Error for training dataset
- Mean Squared Error for test dataset
- R-Squared for training dataset
- R-Squared for test dataset
- Correlation between observed variable and predicted variable
- Twenty percent band of predicted variable

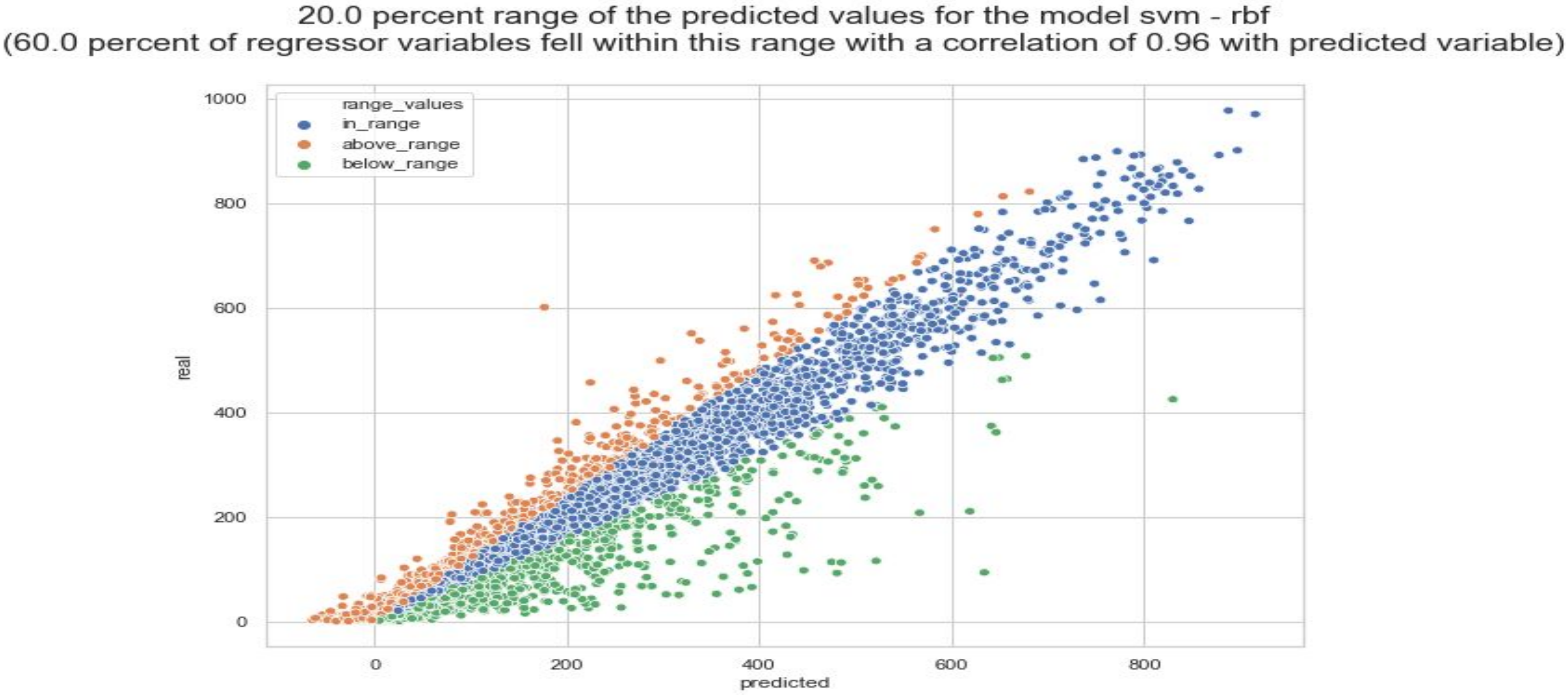
Evaluation Metrics

Algorithms	MSE-TR	MSE-TS	R2-TR	R2-TS	r-corr	Bound-20%
Linear	10,616	10,731	0.68	0.67	0.82	28
Polynomial	2,631	3,047	0.92	0.90	0.95	50
Ridge/poly	2,626	3,024	0.92	0.91	0.95	50
Randomforest	604	2,990	0.98	0.90	0.95	54
SVM	1,906	2,337	0.94	0.93	0.96	60

Mean Squared Errors



Band for Predicted Variable



Recommendation

Based on our analysis of algorithms, we recommend support vector machine for production as this model performed very well in all metrics, especially shows a right balance between overfitting and variance. The second choice falls on polynomial regression which also promises good performances.

END