Predictive Analysis of Seoul Bike Sharing Demand

Myeongin (David) Wang, Prof. Matthew A. Lanham
Purdue University, Daniels School of Business
wang4918@purdue.edu

ABSTRACT

In this project, I developed a linear model to predict the number of Seoul's bike-sharing demand. The goal was to identify whether the system would function on a given day by achieving better resource allocation, maintenance planning, and customer service. The prediction is critical for operational efficiency, cost management, and customer satisfaction. The observation includes 365 days from December 2017 to November 2018. The dataset contains various features including, 'rented bike counts', 'hour', 'temperature', 'humidity', etc. Python was the main programming language for this project. Libraries such as Pandas, NumPy, Matplotlib, and Seaborn were used for data analysis and visualization purposes. Three models – multiple linear regression, ridge regression, and the Lasso - were conducted with 8760 observations and evaluated using 5-fold cross-validation. Model performance was evaluated based on the adjusted R-squared value.

BUSINESS PROBLEM

Urban mobility plays a crucial role in shaping the livability and sustainability of cities worldwide. As cities continue to grow, the demand for efficient and sustainable transportation solutions increases. Bike-sharing systems offer healthier lifestyles, and solutions to alleviate traffic congestion and reduce air pollution. According to McKinsey, the future of mobility is evolving rapidly, with micro-mobility solutions like bike-sharing gaining momentum as urban populations seek more convenient and eco-friendly transportation options (McKinsey, 2023). Additionally, Gensler highlights the transformative impact of micromobility on cities, emphasizing its potential to enhance urban mobility and redefine urban landscapes (Gensler, 2022).

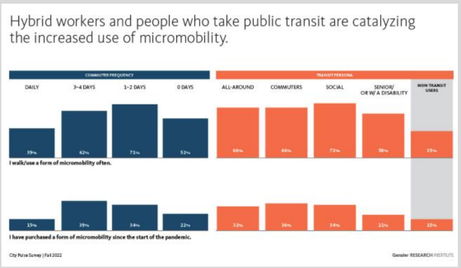


Fig 1. Response to the usage of micro-mobility

Source:
- What the Rise of Micromobility Means for the Future of Cities December 15, 2022 | By Sofia Song, Stella Donovan
https://www.gensler.com/blog/what-micromobility-means-for-the-future-of-cities
-The future of mobility: Mobility evolves
https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-future-of-mobility-mobility-evolves

ANALYTICS PROBLEM FRAMING

- ✓ The purpose of the model is to accurately predict whether the individual used the bike-sharing service based on the given features on the dataset, where the target variable is "rented bike counts".
- ✓ I used the adjusted R-squared value to measure the performance of each model, which measures the percentage of variance in the dependent variable explained by the independent variable.

RESEARCH QUESTIONS

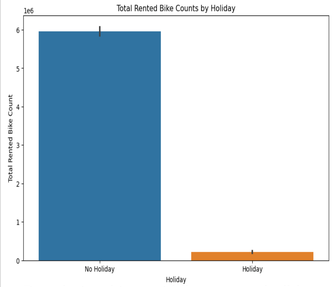
- ✓ How accurately can we predict the operational status of Seoul's bike-sharing system using machine learning?
- ✓ Which method is the best model to predict the demand of Seoul's bike sharing service?

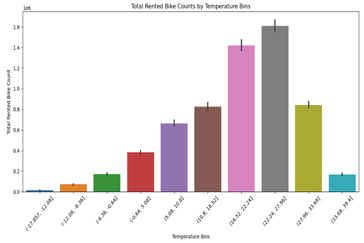
DATA

The dataset is from the UC Irvine Machine Learning Repository, https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand. The chosen dataset contains 8760 observations with no missing values.

Date	Rented Bike	Hour	Temperature	Humidity	Wind Speed	Visibility	Dew Point Temperature	Solar Radiation	Rainfall	Snowfall	Seasons	Holiday	Functioning Day
Date	Integer	Integer	Continuous	Integer	Continuous	Integer	Continuous	Continuous	Integer	Integer	Categorical	Binary	Binary

Feature engineering: remove the column date, which shows the past date record that is not useful in model building
Encoding: encoding categorical variables using one-hot encoding
Handle missing value: since there are no missing values, skip the imputation step
Validation set design: Used 5-fold cross-validation
Standardization: standardize the features

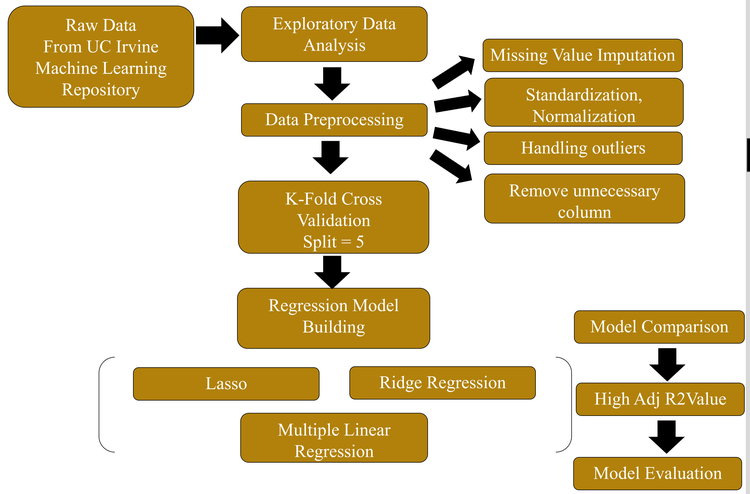




People tend to rent bikes on no holidays. Most of the bikes were rented on days with temperature between 22 and 27 Celsius. It is possible to infer that they are using shared bikes for commuting during the warm temperatures.

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

METHODOLOGY



Flow chart of EDA Process

MODEL BUILDING AND EVALUATION – STATISTCIAL PERFORMANCE

For this predictive analysis, I used multiple linear regression, Ridge regression, and Lasso to predict the outcome
The adjusted R-squared value of each model is:

- Multiple Linear Regression: 0.6518 (+/- 0.0184)
- Ridge Regression: 0.6518 (+/- 0.0184)
- Lasso: 0.6489 (+/- 0.0170)
- Based on the given results, multiple linear regression and ridge regression have identical adjusted R-squared values of 0.6518. This means the model can explain 65.18% of the dependent variable which is rented bike counts.
- Either of these methods is the best model to predict the number of rented bikes in Seoul.
- Since ridge regression includes regularization, it is more robust in various scenarios.

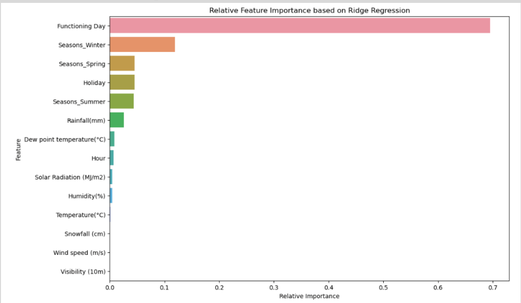
```
Linear Regression:
  Adjusted R^2: 0.6518 (+/- 0.0184)
  MAE: 5.6141 (+/- 0.0766)
  MSE: 53.8389 (+/- 2.2439)
  RMSE: 7.3359 (+/- 0.1533)

Ridge Regression:
  Adjusted R^2: 0.6518 (+/- 0.0184)
  MAE: 5.6144 (+/- 0.0764)
  MSE: 53.8356 (+/- 2.2392)
  RMSE: 7.3357 (+/- 0.1530)

Lasso Regression:
  Adjusted R^2: 0.6489 (+/- 0.0170)
  MAE: 5.6694 (+/- 0.0629)
  MSE: 54.2937 (+/- 1.9542)
  RMSE: 7.3672 (+/- 0.1327)
```

MODEL EVALUATION – BUSINESS IMPLICATIONS

Since there are numerous independent variables that affect the target variable, 0.6185 can be considered to be a reliable adjusted R-squared value. The high performance of the ridge regression allows the business to make strategic decisions with confidence, reducing uncertainty and improving overall service reliability.

With accurate prediction, Seoul Bike can improve in the following areas:

- ➢ Customer Service Notifications: Develop a precise notification system. It can send notifications to customers about the system's availability, enhancing user experience.
- ➢ Service Reliability: Ensure customers can rely on the availability predictions, improving satisfaction and retention.
- ➢ Operational Efficiency: Optimize staffing levels and resource distribution to avoid underutilization or shortages.



CONCLUSIONS

Accurate prediction for bike-sharing system usage is crucial for efficient resource allocation and improving user satisfaction. The ultimate goal is to promote sustainable urban transportation.

- ➢ Both multiple linear regression and ridge regression have the same adjusted R-squared value. However, ridge regression is preferred as the method is more robust in many situations.
- ➢ Ridge regression is the best model to predict whether the individual used the bike-sharing service in Seoul. The model has an adjusted R-squared value that explains 65.18% of the dependent variable (rented bike counts).
- ➢ As the dataset does not consider unexpected disruptions such as special events and infrastructure changes, the dataset may not include all the possible factors.