Bike Sharing Demand

Introduction

Though the business of bike sharing/rental has existed ever since the first bicycle was made available to the public, only now has the venture gone mainstream. Thanks to technological and logistical advances, sharing/renting a bike has become easier and more accessible than ever. Currently, there are over 500 bike-sharing programs around the world. Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. The data generated about the duration of travel, departure location, arrival location, and time elapsed of a bike rented by these systems are explicitly recorded.

As bike-sharing gets more popular, it is important for rental companies to better analyse the data collected to provide better services and in turn minimise losses associated with running such a high-investment/low-margin business. To demonstrate the value of such insights, we are using the data from the Capital Bikeshare program in Washington, D.C..

Objective is to predict the hourly demand of bikes based on historical data.

Dataset

We have 2 years of rental data which provides information about datetime, season, holiday, weather, and others of the bike rented. We have divided our dataset into training and test dataset to perform statistical modelling. The training set contains data of the first 19 days of each month and the test set has data from the 20th to end of the month.

Exploratory Data Analysis (EDA)

Numerical Variables

We plotted a correlation matrix for all the numerical variables to understand the relationship of each predictor with the response and other predictors.
Observations:

- Temp and Atemp have high positive correlation with the response and each other. Hence, we will only choose one of these variables as keeping both will introduce multicollinearity.
- Casual and Registered are highly correlated with bike count as these columns are split of total bike count depicting non-registered vs registered user rentals.

Categorical Variables

There are 5 categorical variables: weather, season, hour of the day, holiday, and working day. To understand the impact of these variables on our response, we are checking the average number of bikes rented per hour for each variable.
Observations:

- Weather, season, and hour of the day have an impact on the target variable.
- Working day and holiday variables have very less impact on our response.

Data Preparation

- To handle the outliers in our target variable we have removed records that fall outside of the 95% confidence interval for the purpose of modelling.
- The distribution of the target variable is right-skewed. To correct the right-skew we have applied logarithmic transformation.
- Using one hot encoding we have converted each categorical variable into a new binary column where the presence of the category is marked with a '1' and the absence with a '0'.

Modelling

We tried linear regression and regression trees on our data to find the best fit for our data:

- **Linear Regression-** Linear model does not seem to be the ideal solution for our dataset as the model barely improves through various iterations. The following models were used:
    - Base Linear Model- Throw in all variables to create a base linear model
    - Lasso and Ridge Regression- This is done for regularisation only
    - 2nd linear model- This was done to predict without using the insignificant variables identified in the above two steps
    - Forward and Backward Regression- To identify the best set of predictors and to check if it helps in better estimation

- **Regression Trees-** We used random forest and boosting models.

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

- Depth = 10
- Number of trees = 1000
- Shrinkage factor = 0.02

Conclusion

- Hour holds the highest importance.
- People are more likely to rent bikes during moderate temperatures and clear weather.
- 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.