[Company name] | [Company address]

Prediction of Ridership

ST 516 – Mid-Term Project

# Executive Summary

In this project, we predict the number of rides for registered and non-registered, casual riders on a given day for the city Washington, DC. The data has been collected for couple of years and the prediction is made out the present dataset. The prediction model used in this project is **k-folds cross validation** because of its high interpretability and moderately high stability in prediction.

**Key Findings:**

1. The ridership has found to have increased over the year by XX % .
2. The ridership was found to be heavier on Sunny days as compared to other days irrespective of the season.

# Introduction

# Data

The dataset “Bike-Sharing-Dataset” was obtained by the UCI Machine Learning Repository. This is a collection of databases, domain theories and data generators which are used by the machine learning community for empirical analyses. This dataset contains the daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. Capital bikeshare has over 350 stations in Washington, D. C. Bike sharing systems are a new way of traditional bike rentals. We have two sets of riders – Registered riders (one’s who have an active membership and registration) and Casual riders (one’s who rent it if they have a need and are not registered).

The dataset can be found in the link - <https://www.capitalbikeshare.com/system-data>.

Since the given dataset is cleaned in our case, we do not require any data massaging. In order to get the initial view on the data, we split the data into two tables – One with Registered drivers and one with casual drivers.

When we plotted the correlation matrix for the above dataset, the following results were obtained.

# Methods

## Linear-Regression Model

In this method of analysis. We have bifurcated the data in 2 parts casual, registered and fit a linear model with 10 predictors (including the categorical variables). Then we generated interaction terms and fit a complete second order linear model with 13 predictors (including the categorical variables) for casual riders and registered riders and then the observation shows that there is negligible difference in the adjusted R2.

## k-folds cross validation

In this model, we took 10-folds cross validation technique to perform the analysis.

## Ridge and Lasso Regression

## PCR Analysis

# Results

# Conclusion