Project Proposal

Intro to statistical computing

Abhijeet Vichare

Predictive modeling in R

Introduction

An open source dataset published on Kaggle will be used for the project. The project would involve regression analysis for prediction. The models used will start with Linear regression and test multiple methods on top of it such as regularization, cross validation, testing out Grid Search/Random Search CV methodologies to find the optimal hyperparameter values.

Methods used in the project.

In addition to regression analysis, multiple steps for data preparation, importing, wrangling, data visualizations, exploratory data analysis will be used to understand the data better and prepare them for the model.

Detailed steps included in the Data exploration phase:

1. Data Wrangling
    a. Data importing
    b. Checking the data types of variables and performing necessary steps to change the data type based on relevancy.
    c. Univariate Statistics of the variables
    d. Dealing with null values
    e. Outlier analysis
2. Data Visualization
    a. Bivariate analysis
    b. Correlation analysis
3. Feature importance
    a. Chi square test
    b. ANOVA test
    c. Correlation test
4. Linear Regression
    a. Linear Regression assumption validation
    b. Model training
    c. Hyperparameter tuning
    d. Regularization
    e. Result visualization