Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Contribution Role :

1. Nikunj Sonule (nikunj.sonule10@gmail.com)
    1. Data Preprocessing
    2. Exploratory Data Analysis
        - Univariate Analysis on numeric features
        - Bivariate Analysis
        - Analysis from categorical variables
    3. Fitting different models
        - Linear Regression
        - Building Lasso, Ridge and ElasticNet Regression
        - Decision Trees
        - Random Forest
    4. Finding best parameters by GridSearchCV in Lasso, Ridge and ElasticNet Regression
    5. Getting Feature Importance on Decision Tree and Random Forest model
    6. Model Evaluation Metrics
        - Mean square error
        - Root mean square error
        - R2
        - Adjusted R2

Please paste the GitHub Repo link.

Github Link:- https://github.com/nikunjsonule/Bike-Sharing-Demand-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Rental Bikes are introduced in many urban cities for the enhancement of mobility comfort. We were provided with ‘Seoul Bike Demand Prediction’ dataset to perform machine learning task, were to get insights from the independent and dependent variable of a dataset. By analyzing the problems with count of bikes rented at each hour the demand is getting increases with some specific times, seasons and whether. We’ve a data from 1st December 2017 till 30th November 2018. In this particular time duration we’ve all record and data with different seasons, whether, temperature, rented bike count, hour, humidity, visibility, solar radiation, rainfall, snowfall, holiday and functioning day.

As the first step, I performed data preprocessing and EDA part. I extracted features from ‘Date’ column, to get month, dayofweek and weekday_weekends. In EDA plotting heatmap for correlation, univariate and bivariate analysis, graphs and other plots to get some insights from the data.

After data preprocessing, our data was ready to fit into the models. Different model used to get best accuracy score, models are – Linear Regression, implementing Lasso, Ridge and ElasticNet Regression, Decision Trees and Random Forest model.

The third step is to do some hyperparamter tuning. Doing GridSearchCV on Regularization(i.e Lasso, Ridge and ElasticNet) part to get best accuracy and to shrink the coefficient. But it didn’t get a useful accuracy.

As to get more insights from data, that which independent variable is more important when we fit data on model. Looking towards feature importance I noticed ‘temperature’ and ‘Hour’ are the most important variable for Decision tree and Random Forest model. There is slightly competition between ‘Solar Radiation’ and ‘Functioning Day_Yes’ with more necessary feature for model. The ‘temperature’ is the most feature importance among all features.

The final section is to observe model evaluation metrics. Linear regression, Lasso, Ridge and ElasticNet regression shows accuracy score with 56%. As for Decision Tree and Random Forest we get 77% and 84% adjusted R2 score respectively.

So accuracy of the best model is 84% for this dataset. This performance could be due to various reasons: no proper pattern of data, it may be some negligible noise in the data, different model give different accuracy score.

That's how I have accomplished my project work in Seoul Bike Sharing Demand Prediction.

Drive link :-

https://drive.google.com/drive/folders/1beCbj-MSMX7hN_8M3m4D10rs2jtqUXNz?usp=sharing