Contents: 1 Introduction

1.1Problem Statement

1.2Data

2 Methodology

2.1 Missing Value Analysis

2.2 Univariate Analysis (Distribution of Numeric variables)

2.3 Bivariate Analysis (Distribution of Categorical variables with Count)

2.4 Bivariate Analysis (Distribution of Continuous variables with count)

2.5 Outlier Analysis

2.6 Feature Selection

3 Modelling

3.1 Model Selection

3.2 Multiple Linear Regression

3.3 Decision Trees

3.4 Random Forest

4 Conclusion

4.1 Model Evaluation

1-identify the problem

2-type fo problem

3-[handle missing value](https://towardsdatascience.com/handling-missing-values-in-machine-learning-part-1-dda69d4f88ca)

**1.Ignore the data row**

**2. Back-fill or forward-fill** to propagate next or previous values respectively:

dataframe.fillna(method='bfill',inplace=True)

Note that the NaN value will remain even after forward filling or back filling if a next

or previous value isn’t available or it is also a NaN value

**3.** **Replace with some constant value outside fixed value range-999,-1 etc**

This method is useful as it gives the possibility to group missing values as a separate category represented by a constant value.

**4.** **Replace with mean, median value using Imputer**

MEAN: Suitable for continuous data without outliers

MEDIAN :Suitable for continuous data with outliers

4-[Dealing with imblance class methode](https://imbalanced-learn.readthedocs.io/en/stable/api.html) - [link-2](https://www.edyoda.com/course/1416?episode_id=461)

Oversampling methode

1.SMOTE

2. ADASYN

Better then SMOTE.genrate data for highly misclassified data.

Undersampling methode

1.RandomUnderSampler

5.[**Anomaly Detection outliear**](https://www.edyoda.com/course/1416?episode_id=432)

For univeratant type data

1.zscore-return standard deviation of data .use For normal distribution type data

2.IRQ-for univariate data that not follow gaussian distribution

[6-Feature Selection Techniques](https://www.edyoda.com/course/1416?episode_id=395)

For univeratant type data

1.varianceThreshold for **unsupervised model**

**Classification**

2.chi-squre caterogy feature

3.ANOVA for continue feature

Regression

1.f\_regression **f value**

**2.for repeated pattern data use Mutual\_info\_regression,mutual\_info\_classification**

**Other methode -**

**1-selectKbest (score\_fun=feature\_selection.f\_classif)**

**2.selectfrommodel (using inside ml model which suppport this methode like LR)**

**3.recursive feature selection**