

Summary

The first process we do with our data-set is Data Inspection. In data inspection we perform **.shape**, **.describe()**, **.info()** methods to inspection.

Then we perform **Data cleaning**.

1. In data cleaning the first thing we do is we replace all the **'Select' values with NAN** values. T
2. Then we check for null value percentage in columns and drop the columns having more than 40% missing values.
3. For the columns having less than 40% missing values we imputed the columns with appropriate values.
4. We also drop rows accordingly.

Here our data cleaning process is finished.

Then we begin with **Exploratory Data Analysis**:

1. In this section we perform **Uni-variate and bi-variate analysis**. And by this process we infer important insights about various variables.
2. In EDA we also look for **outliers**.

Here our EDA part is done.

In **Data preparation** part we perform:

1. In data preparation we firstly **convert binary variables to 1 and 0**.
2. After this we perform **dummy encoding** on categorical variables.
3. After creating Dummy variables we concat dummy variables with the actual data-set and drop the categorical columns on which we performed dummy encoding.

4. Now we **split our data-set into X_train, X_test, y_train, y_test.**
 5. After splitting our data-set into train and test, we perform normalization scaling, that is **MinMax scaling.**
- Here our Data preparation part is finished. Now we'll look into Feature Selection.

In **Feature Selection** we perform:

1. We build a model using **SK_learn and perform RFE** on that. RFE stands for Recursive Feature Elimination.
2. After performing RFE we look for the columns that are selected by the RFE algorithm.
3. We will be using these same columns for further model building purposes.
4. And also we will perform manual feature elimination from this same set of columns.

Here we are done with the Feature Selection. Now we will head towards model building.

In **Model building** we perform:

1. Firstly we create a model using **statsmodels.**
2. After creating this model we look for its detailed summary using **.summary().**
3. After going through summary we will find **VIF.**
4. VIF stands for Variance Inflation Factor.
5. Now according to the values of **P-value, which is present in statsmodels summary, and VIF we will perform manual feature elimination.**
6. In manual feature elimination we look for these stats and then accordingly remove the columns, then we again build the model and then again remove the columns, and this process goes on until we get desired result.

After creating a suitable model, we make **predictions on train data-set** itself:

1. We compare the actual dependent variable with the predicted dependent variable.
2. We have various metrics on which we judge our model on, such as **accuracy, sensitivity-specificity, precision-recall**.
3. According to these metrics we check efficiency of our model.
4. Also according to these metrics we find the Optimal cut-off
5. And according to this **optimal cutoff**, we make the decision whether to consider it as 1 or 0.

After finding the optimal cut-off, we perform a **model evaluation on Test data-set**:

1. Firstly we predict the dependent variable values.
2. Then we **compare the actual y_test values and the predicted y_test values**.
3. And according to this comparison we judge how good our model is.
4. We use the same metrics that we used to judge our model on train data-set on our test data-set as well.
6. We use accuracy, confusion matrix, sensitivity-specificity, precision-recall, etc to test our model.

After our model evaluation part we **make a list of all the important variables**. And we use these variables to infer important insights.

In the end we also provide **business recommendations** for the problem statement based on our detailed analysis.