### **Credit-Card-fraud-detect-under-over-sampling**

1. Import all the libraries required for this project (Pandas,numpy,matplotlib,seaborn,sklearn,etc.)
2. All the Datasets Regarding this Case-Study has been Imported.
3. Perform EDA-Clean and Transform the data
   * 1. Missing value Treatment.
     2. Remove duplicate values.
     3. Outlier capping.
     4. Data Transform
4. Train Dataset has been separated from Test Dataset where Train dataset contains 2,20,530 rows and Test dataset contains 55,133 rows.
5. Train Dataset’s Info and Description is Analyzed.
6. Handling Imbalance Dataset use method -UnderSampling and OverSampling
7. It is Required to transform all the chosen X-variables and Y-variable, else the prediction cannot be accurate.
8. For undersampling Train Dataset has been separated from Test Dataset where Train dataset contains 756 rows and Test dataset contains 190 rows.
9. Train Dataset has been separated from Test Dataset where Train dataset contains 4,40,304 rows and Test dataset contains 1,10,076 rows.

Algorithms chosen for this Case-Study are:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier

**UnderSampling-**

* **Logistic Regression Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Logistic Regression Models** | **Accuracy Score** | **Precision** | **Recall** | **F1-score** |
|  | **93.684%** | **96.875%** | **91.176%** | **93.939%** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decision Tree Classifier Models** | **Accuracy Score** | **Precision** | **Recall** | **F1-score** |
|  | **90.000%** | **91.919%** | **89.215%** | **90.547%** |

* **Decision Tree Classifier Model:**
* **Random Forest Classifier Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest Classifier Models** | **Accuracy Score** | **Precision** | **Recall** | **F1-score** |
|  | **93.157%** | **96.842%** | **90.196%** | **93.107%** |

**OverSampling-**

* **Logistic Regression Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Logistic Regression Models** | **Accuracy Score** | **Precision** | **Recall** | **F1-score** |
|  | **94.758%** | **97.300%** | **91.693%** | **94.413%** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decision Tree Classifier Models** | **Accuracy Score** | **Precision** | **Recall** | **F1-score** |
|  | **99.814%** | **99.747%** | **99.881%** | **99.814%** |

* **Decision Tree Classifier Model:**
* **Random Forest Classifier Model:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest Classifier Models** | **Accuracy Score** | **Precision** | **Recall** | **F1-score** |
|  | **99.992%** | **99.985%** | **100%** | **99.992%** |

**To Conclude this Report:** 

**Bar chart represent accuracy scores of different models in undersampling**



**Bar chart represent accuracy scores of different models in Oversampling .**

* **Above Charts shows us that in undersampling, we can use Logistic Regression Model performed much better .**
* **In oversampling , we can use Random Forest Classifier Model performed much better**
* **We can also improve on this accuracy by increasing the sample size.**
* **When comparing error precision & recall for 3 models , the Logistic Regression performed much better than other models in underSampling and in oversampling Random Forest Classifier perform better**