Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) **Arbaaz Malik :** **Email:** malikarbaaz267@gmail.com

- EDA
  a) Dropping duplicates
  b) Handling null and missing values
  c) Handling Outliers

- Feature Creation
  a) Renamed all features with suitable meanings
  b) Made new column out of month wise payment

- Univariate and Bivariate Analysis
  a) Dependent variable
  b) Limit of balance
  c) Marriage plot, Education plot and Sex plot.
  d) Defaulter status with sex, education and marriage.
  e) Total number of customers with credit limit

- Feature encoding
  a) One hot encoding
  b) Drop unwanted columns

- Correlation analysis between independent/dependent variables.
- As we have imbalance data, so handled imbalance data using random over sampler.
- Machine Learning Classification algorithms :
  a) Logistic Regression
  b) Stochastic Gradient Decent
  c) Decision Trees Classifier
  d) Support Vector Machine

2) **Huzaifa Khan :** **Email:** huzaifakhan2974@gmail.com

- EDA
  a) Dropping duplicates
  b) Handling Null/nan values
  c) Handling Outliers

- Feature Creation
  a) Renamed all features with suitable meanings
  b) Replaced low count column values with other values

- Univariate and Bivariate Analysis
  a) Normalized default payment and plotted the graph
  b) Sex ratio

c) Education
d) Payment done in Aug and July
e) Pair plot of previous payment of each month
f) Defaulter status with sex, education and marriage
g) Month wise payment according to Defaulter status

- Feature encoding

a. One hot encoding
b. B. Dropping unwanted columns

- Correlation analysis between dependent/independent variables
- Splitting dependent and independent features
- Train test split on dependent and independent features
- Handled imbalance data using random over sampling method
- ML Regression algorithms used:
  a) Logistic regression
  b) Decision tree classifier
  c) XGBOOST classifier

Please paste the GitHub Repo link.

Github Link:- https://github.com/Malikarbaaz/CREDIT_CARD_DEFAULT_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)

As we know in todays world no one is without a credit card , so it has become a challenge for the companies to keep record of all customers who make late payments and who makes it on time to decide whether their limits should be increased or not, so here with the help of this dataset we will try and find the customers who are defaulters.

This project is aimed at predicting the case of customers' default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

Modelling:-

We will build Five models

- **Logistic Regression**
- **Stochastic Gradient Descent**
- **Decision Tree Classifier**
- **Default XGBoost Classifier**
- **Support Vector Machine**

Performance Metrics

- **Precision is a good metric to use when the costs of false positive(FP) is high.**

  Precision = TP / (TP + FP)

- **Recall is a good metric to use when the cost associated with false negative(FN) is high.**

  Recall = TP / (TP + FN)

- **F1-score is a weighted average of precision and recall. Thus, it considers FP and FN. This metric is very useful when we have uneven class distribution, as it seeks a balance between precision and recall.**

  F1-score = 2 (precision recall) / (precision + recall)

conclusion:

1. When we load the data we check for duplicate and null values if there but luckly we see that there is no null values nor duplicate values, still we apply dropna for the betterment of our convinience.

2. Next we rename the column for each features for better understanding.

3. From above graph we check the distribution of defaulter vs non defaulter and we see around 78% are non defaulter and 22% are defaulter. Also we check for Marriage, Education, Sex with respect to defaulter and we found in marriage more number of defaulter is Female, in Education more no. of defaulter is university and in Marriage more no. of defaulter is single.

4. Using boxplot to detect the outliers and we see that there are so many outliers in the data so we apply IQR(Inter Quartile Range) which is one of the technique to remove outliers.

5. Correlation matrix gives the graphical representation between all the variables and with the help of correlation matrix we see that age and marriage are highly negatively correlated to each other.

6. After that we build the Five models Logistic Regression, Stochastic gradient descent, Decision Tree, Default XGBoost Classifier & Support vector machine and inspite of all the models,the best accuracy is obtained from the Default XGBoost Classifier.

- Using a Logistic Regression classifier, we can predict with 69.38% accuracy, whether a customer is likely to default next month.
- With Stochastic Gradient Descent classifier, 58.84% customer is likely to default next month.
- With Decision Tree classifier, 71.88% customer is likely to default next month.
- With Default XGBoost Classifier, 76.33% customer is likely to default next month.
- And with Support Vector Machine classifier, 77.63% customer is likely to default next month.

7. From above table Default XGBoost Classifier and Support Vector Machine are giving us the best Precision, F1-score, and ROC Score among other algorithms. We can conclude that these two algorithms are the best to predict whether the credit card is default or not default according to our analysis.