Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name : GOPAL JI GUPTA **Email: gopaljigupta45@gmail.com**

Role :-

1) Data Cleaning :-
   - Dealing with null values, duplicate data and outliers present in our data.
2) Exploratory Data Analysis :-
   - Plotting the dependent variable and univariate and bivariate analysis of dependent and independent variables.
   - Checking and visualizing the correlation between our dependent and independent variables.
   - Visualizing the relationship between several pair of our variables.
3) Data Preprocessing & Feature Engineering :-
   - Checking for and Dealing with multicollinearity present in our dataset.
   - Performing encoding for our categorical data.
   - Scaling the data and splitting it into train and test sets.
4) Model Implementation :-
   - Fitting various models on our data and optimizing them via cross-validation.
   - Using these models to make predictions on test and train data.

     The Models implemented are :-
     1. Logistic Regression
     2. Random Forest
     3. K Nearest Neighbor Classifier
     4. Support Vector Classifier
5) Data Visualization :-
   - Using several kinds of charts like pie chart, bar chart, heatmap, pair plot, countplot, boxplot etc to better visualize data and understand correlation, trends and relationship amongst variables.
6) Model performance comparison :-
   - Comparison of all implemented models using various Classification evaluation metrics like Accuracy, Precision, Recall, F1 score, ROC-AUC score, classification report etc.
7) Conclusion :-
   - Drawing some insights from the data and the predictions made by our various predictive models on unseen (test) data.

Please paste the GitHub Repo link.

Github Link:- https://github.com/gopaljigupta45/CREDIT_CARD_DEFAULT_PREDICTION

Drive folder link :-
https://drive.google.com/drive/folders/1Nh8BncO5hEPsPoswbdJca-cyZIfJxRgZ?usp=sharing

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

Problem statement :-

Today's world is interconnected. With the rise of technological advancements, every sector has made improvements. In financial sector, this advancement has come in the form of a credit card. This purchase now and pay later strategy enabling tool has been a boon for banks as they have found another method to profit while providing their clients a better service at the same time
.
However, it has its challenges like any other advancement. The challenge for the credit card companies is to find the right people to issue these cards to and keep it out of others' hands. The thing is whenever a client uses a credit card, they are effectively borrowing money. Some people however will not be able to pay the amount and interest on it. So the recovery of that amount is what poses a challenge. So the banks / credit card companies have to be very careful in deciding who they issue a credit card to. So they have to come up with optimal parameters to filter through potential applicants to make sure they avoid potential defaulters lest their money will be hard to recover. This is achieved by learning the characteristics and behavior of defaulters through looking at past data.

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.

Our approach is to fit a machine learning model on this past data and try to predict the closing price for new unseen data using the parameters learned during training.
This way, we can get our model to learn the trends present in the data during training and use that information during prediction.
We will apply various Classification Models for this task such as : Logistic Regression, Random Forest, K Nearest Neighbor and Support Vector classifier.

Conclusions :-

- Most of the credit card users are Female and have higher number of defaults.
- Most of the credit card users are highly educated.
- Single users have more no. of credit cards.
- The number of credit card users goes down with increase in age as old people have less consumption and may not be able to use credit cards and their purchases are usually made by younger family members.
- Using a Logistic Regression classifier, we can predict an accuracy of 67.7% and ROC_AUC score of 0.663
- Using Random Forest Classifier, we can predict an accuracy of around 87.6% and ROC_AUC score of 0.837
- Using K-Neighbor Classifier, we can predict an accuracy of 85.69% and ROC_AUC score of 0.858
- Using Support Vector Machine Classifier, we can predict an accuracy of 76.66% and ROC_AUC score of 0.723
- Random Forest Classifier performs best among all models.
- Logistic Regression is not giving good precision score
- Our best models are Random Forest, K-Neighbor Classifier and Support Vector Machine that score really well on Precision, Recall, ROC_AUC and F1 score with K NEAREST CLASSIFIER & RANDOM FOREST being the best performers as they score the best on nearly every metric.

.....