CREDIT CARD DEFAULT PREDICTION

BY-CHARAN

POINTS TO DISCUSS

DATA SUMMARY

DATA CLEANING

PROBLEM STATEMENT

EDA

FEATURE SELECTION

FEATURE ENGINEERING

DATASET FOR MODELING

APPLYING MODEL

MODEL VALIDATION & SELECTION

CONCLUSION

PROBLEM STATEMENT

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. WE HAVE TO EVALUATE WHICH CUSTOMERS WILL DEFAULT ON THEIR CREDIT CARD PAYMENTS.

FINANCIAL THREATS ARE DISPLAYING A TREND ABOUT THE CREDIT RISK OF COMMERCIAL BANKS AS THE INCREDIBLE IMPROVEMENT IN THE FINANCIAL INDUSTRY HAS ARISEN. IN THIS WAY, ONE OF THE BIGGEST THREATS FACED BY COMMERCIAL BANKS IS THE RISK PREDICTION OF CREDIT CLIENTS. TO ANALYZE AND PREDICT THE ABOVE GIVEN DATABASE, THE CURRENT PROJECT IS DEVELOPED. THIS PROJECT IS AN ATTEMPT TO IDENTIFY CREDIT CARD CUSTOMERS WHO ARE MORE LIKELY TO DEFAULT IN THE COMING MONTH.

STEPS INVOLVED IN SUPERVISED ML MODEL

1. Defining the problem statement
2. Pre-processing the data
3. Splitting the data into train and test data
4. Training the model
5. Evaluating the model
6. Improving the model
7. Deploying the model and monitoring in real-time

DATA SUMMARY /CLEANING

SHAPE OF THE DATA
ROWS=30000
COLUMNS=25

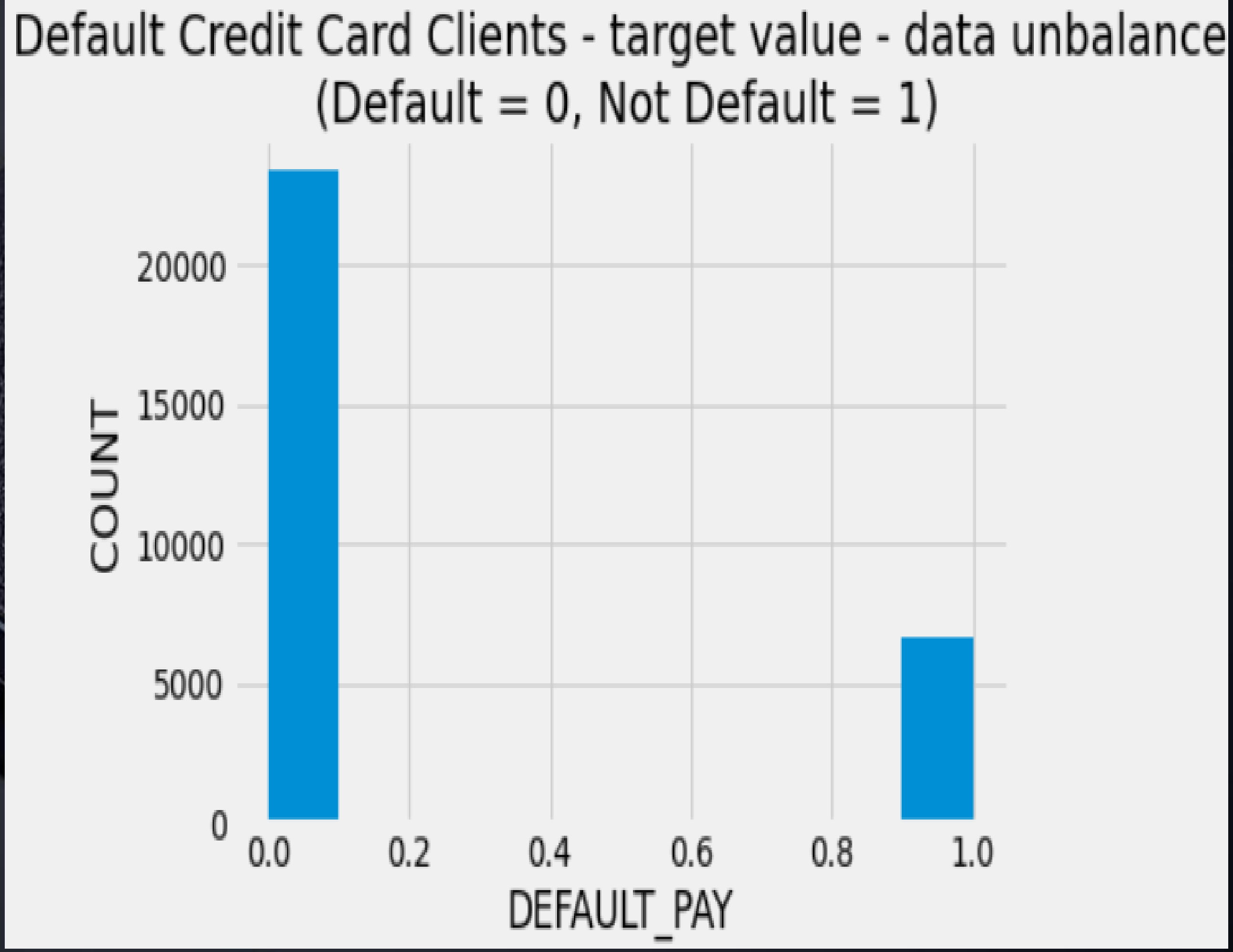
DROP_COLUMN
DROPING UNWANTED
COLUMNS

data
FULLY CLEANED
DATASET

EDA(EXPLORATORY DATA ANALYSIS):

Analysis of different variables:

1.PAYMENT METHOD

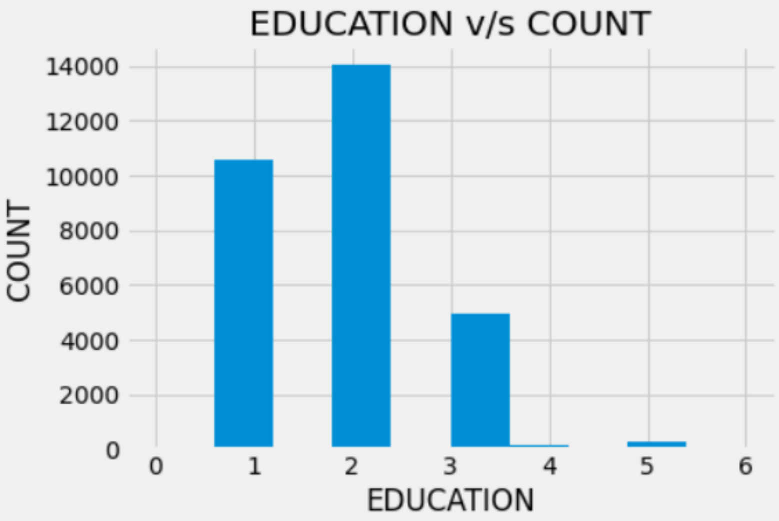


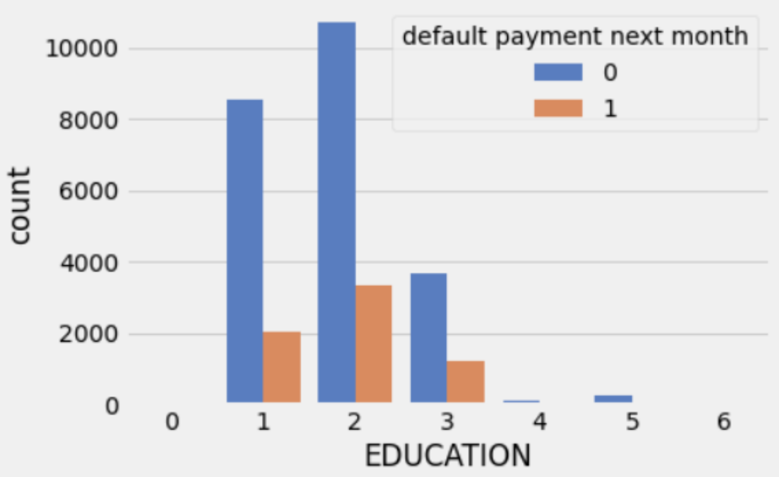
FROM THIS BAR GRAPH ON X AXIS 0 INDICATES A NOT A DEFAULT PAYMENT AND 1 INDICATES THE DEFAULT PAYMENT.

BY LOOKING AT THIS WE CAN CLEARLY SAY THAT DEFAULT PAYMENTS ARE LESS(DEFAULT=1)

2.EDUCATION

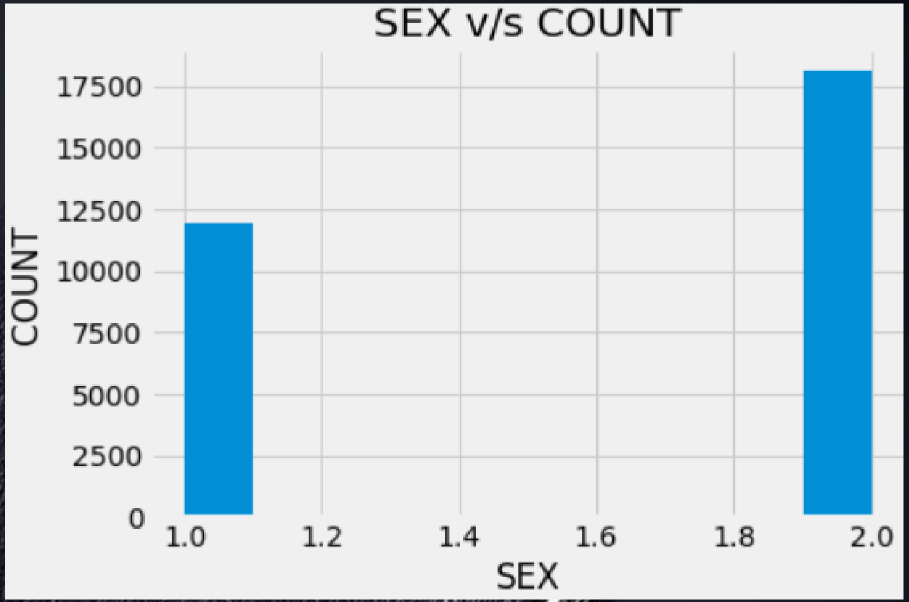
ANALYZING THE CUSTOMERS BASED ON THEIR EDUCATION (1 = GRADUATE SCHOOL; 2 = UNIVERSITY; 3 = HIGH SCHOOL; 4 = OTHERS) IT MIGHT BE USEFUL TO SEE WHETHER THE EDUCATION LEVEL OF THE CUSTOMER IS IN ANY WAY RELATED TO HIS/HER PROBABILITY OF DEFAULT. THE DISTRIBUTION OF DEFAULTS BASED ON EDUCATION LEVEL WILL BE AN INTERESTING CHART TO LOOK AT. FROM THE ABOVE WE CAN SAY THAT MOST OF THE PEOPLE ARE UNIVERSITY EDUCATED FOLLOWED BY GRADUATED SCHOOL MORE NUMBER OF CREDIT HOLDERS ARE UNIVERSITY STUDENTS FOLLOWED BY GRADUATES AND THEN HIGH SCHOOL STUDENTS.

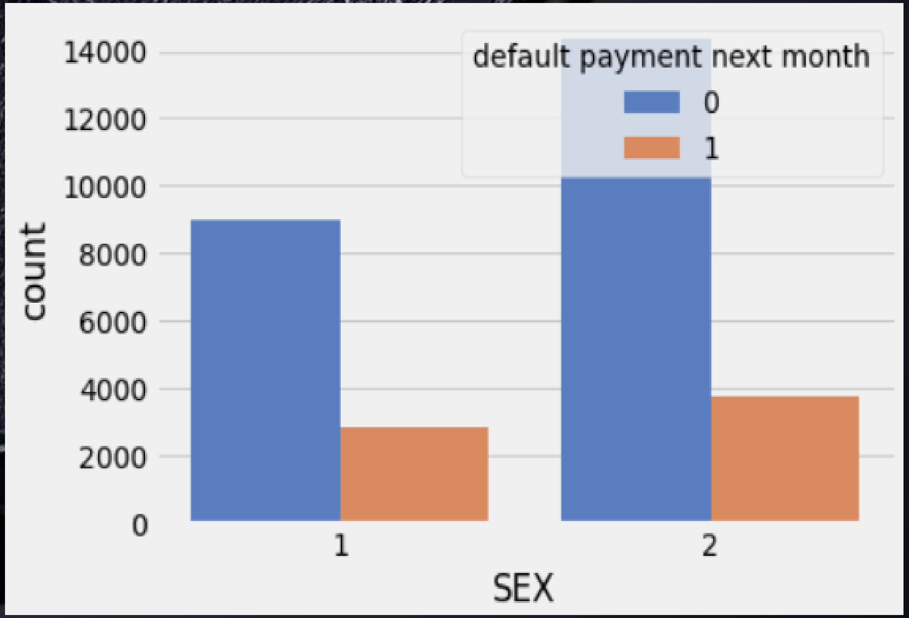




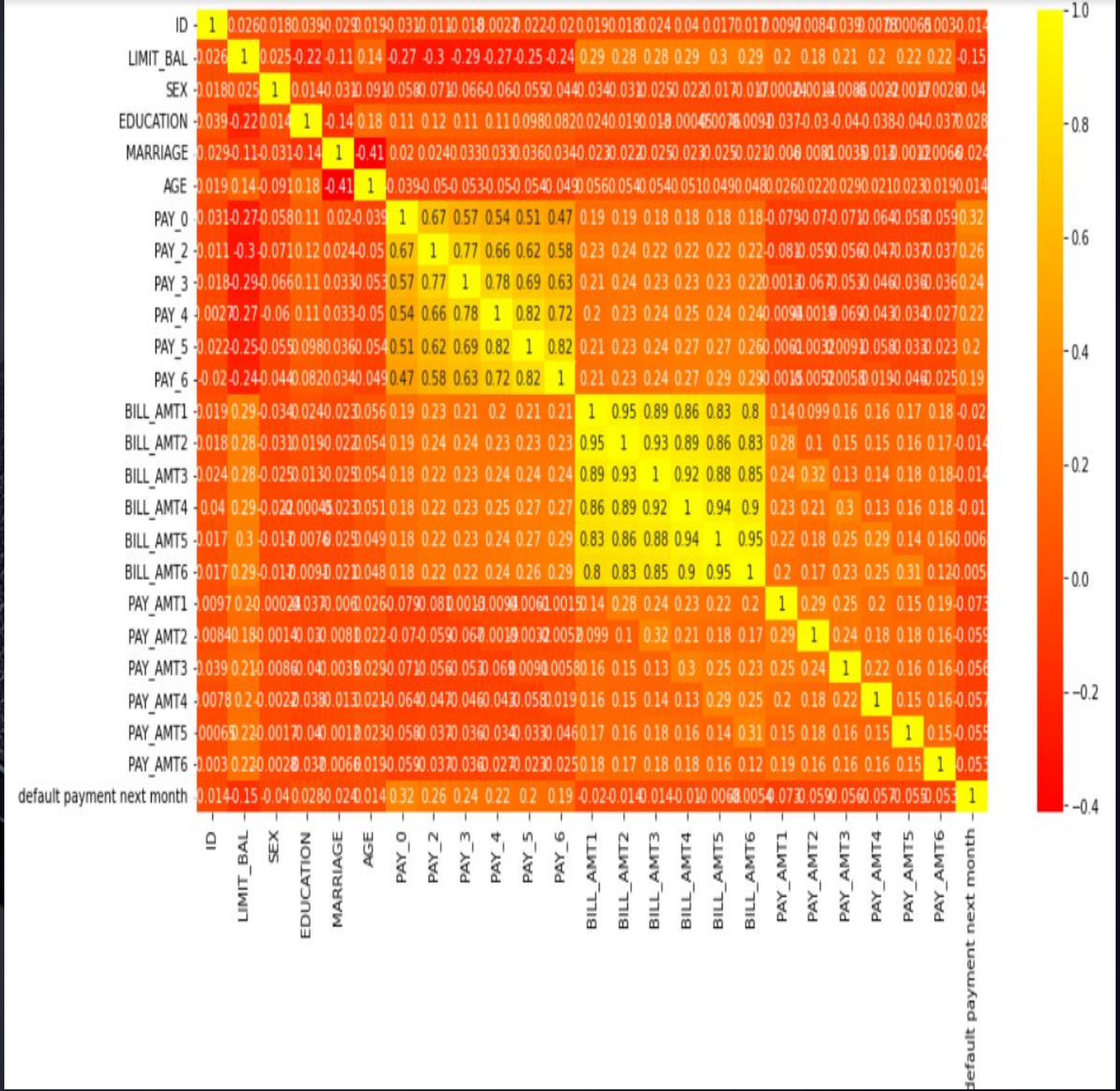
3.SEX:

GENDER (1 = MALE; 2 = FEMALE) IT MIGHT BE USEFUL TO SEE WHETHER THE GENDER OF THE CUSTOMER IS IN ANY WAY RELATED TO HIS/HER PROBABILITY OF DEFAULT.





CORRELATION

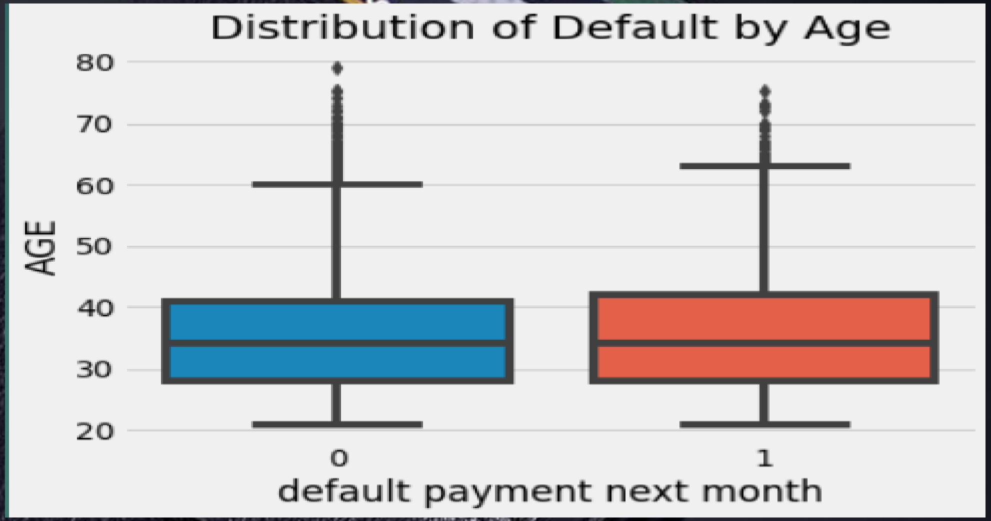


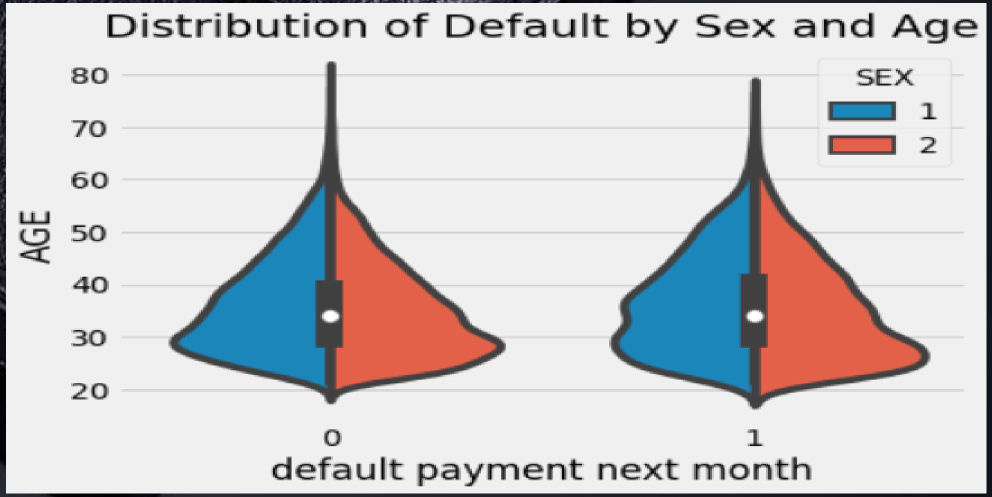
The above figure is an indication about the changes between two variables.

We can plot correlation matrices to know which variable is having a high or low correlation with respect to another variable.

ANALYZING THE AGE OF THE CUSTOMERS

AGE (YEAR) IT MIGHT BE USEFUL TO SEE WHETHER THE GENDER OF THE CUSTOMER IS IN ANY WAY RELATED TO HIS/HER PROBABILITY OF DEFAULT. FROM THE ABOVE GRAPHS WE CAN SAY THAT MOST OF THE CUSTOMERS ARE FROM THE AGE GROUP OF 25-30 YEARS.





TRAINING THE MODEL

FITTING DIFFERENT MODELS

FOR MODELLING WE TRIED VARIOUS CLASSIFICATION ALGORITHMS LIKE:

LOGISTIC REGRESSION

SVC

LOGISTIC REGRESSION USING LLE

SVC WITH LLE

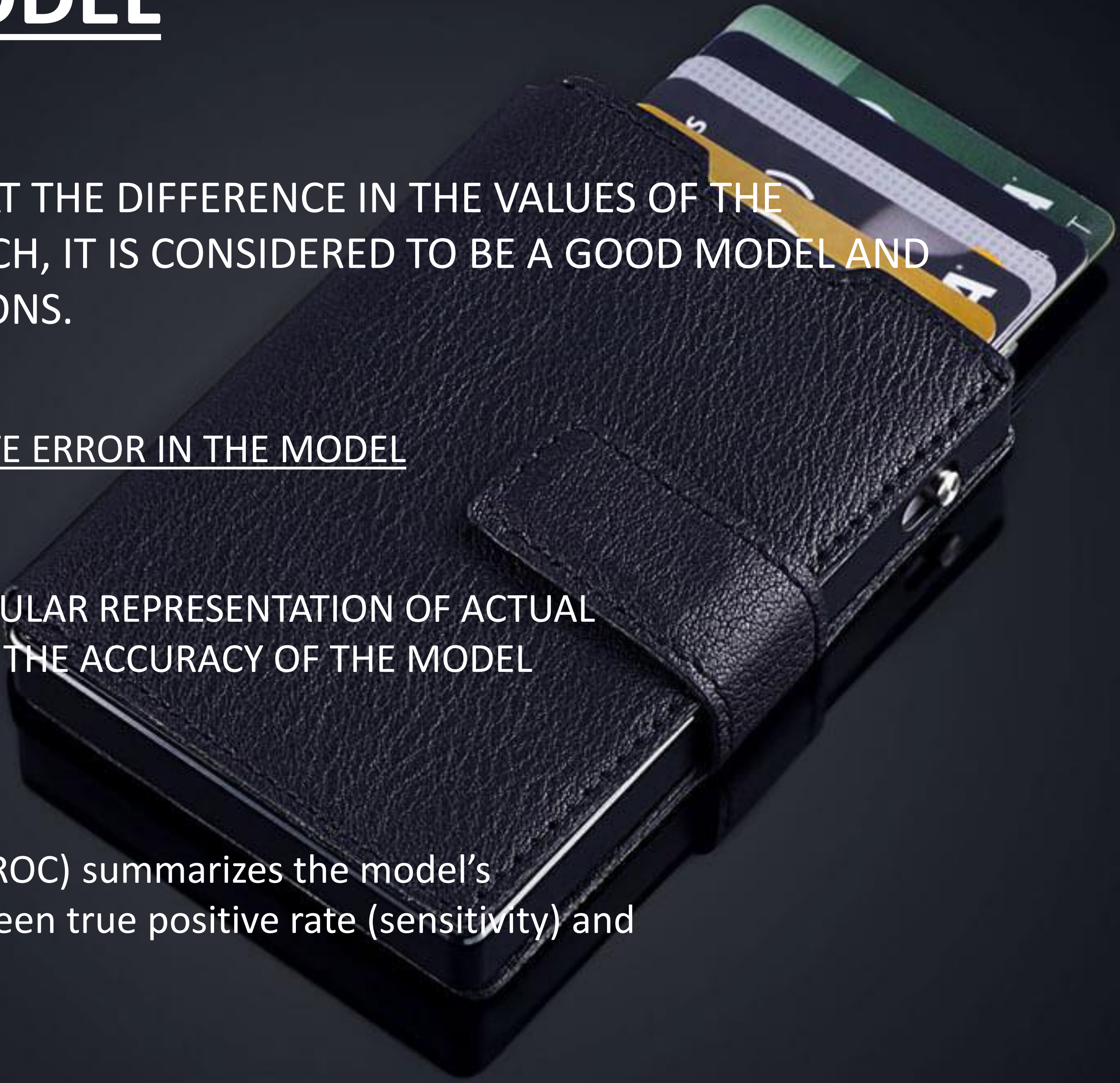
FINE TUNING

ENSEMBLE LEARNING

BAGGING CLASSIFIER

VOTING CLASSIFIER

EVALUATING THE MODEL



AFTER THE MODEL IS BUILT, IF WE SEE THAT THE DIFFERENCE IN THE VALUES OF THE PREDICTED AND ACTUAL DATA IS NOT MUCH, IT IS CONSIDERED TO BE A GOOD MODEL AND CAN BE USED TO MAKE FUTURE PREDICTIONS.

FEW METRIC TOOLS WE CAN USE TO CALCULATE ERROR IN THE MODEL

CONFUSION MATRIX: IT IS NOTHING BUT A TABULAR REPRESENTATION OF ACTUAL VS PREDICTED VALUES. THIS HELPS US TO FIND THE ACCURACY OF THE MODEL AND AVOID OVERFITTING.

ROC Curve: Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade-offs between true positive rate (sensitivity) and false positive rate(1- specificity).

ML MODEL ACCURACY

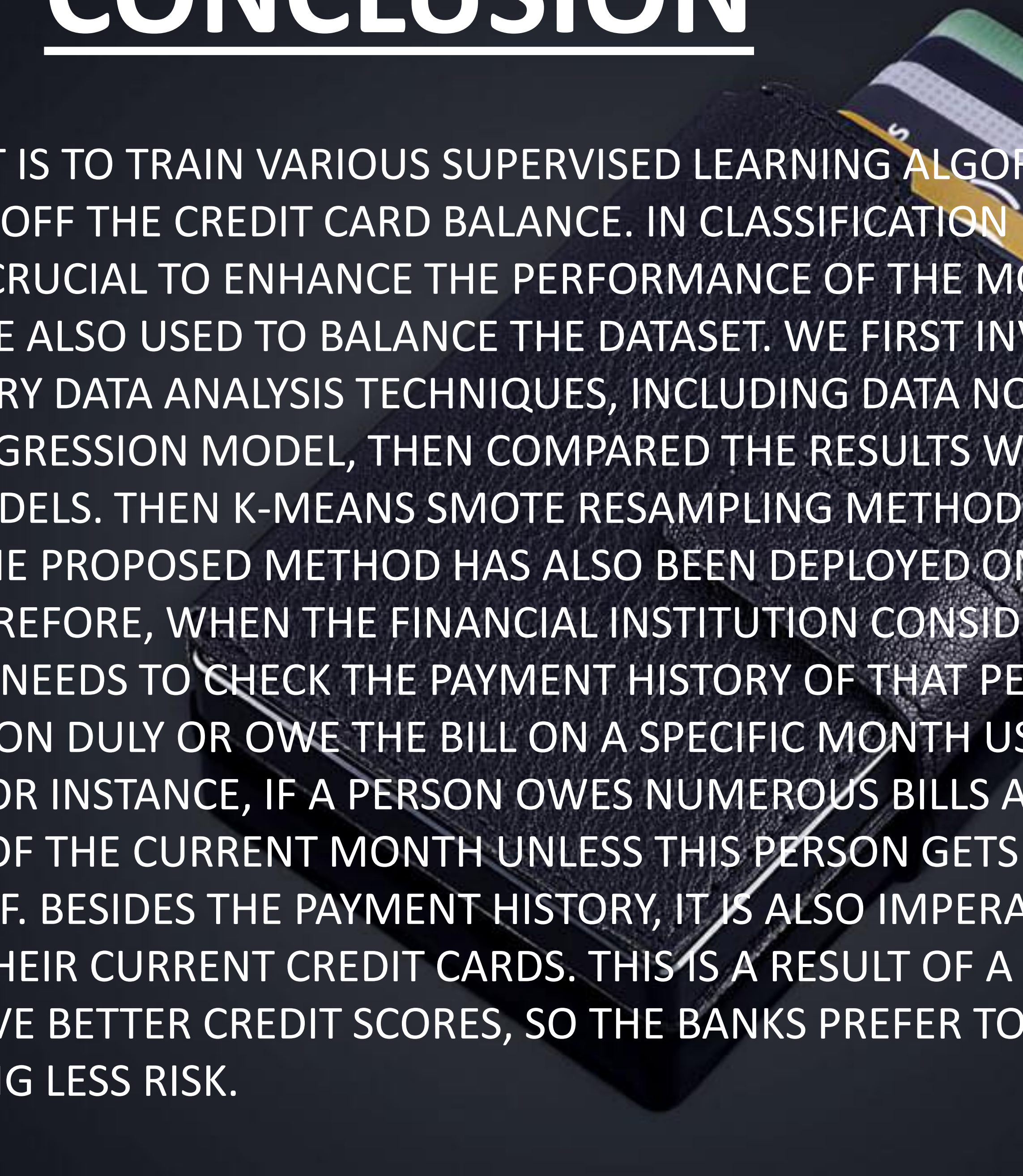
Accuracy of Logistic Regression Model: 0.8206

Accuracy of SVC Model: 0.8198

Accuracy of Bagging Classifier: 0.8206

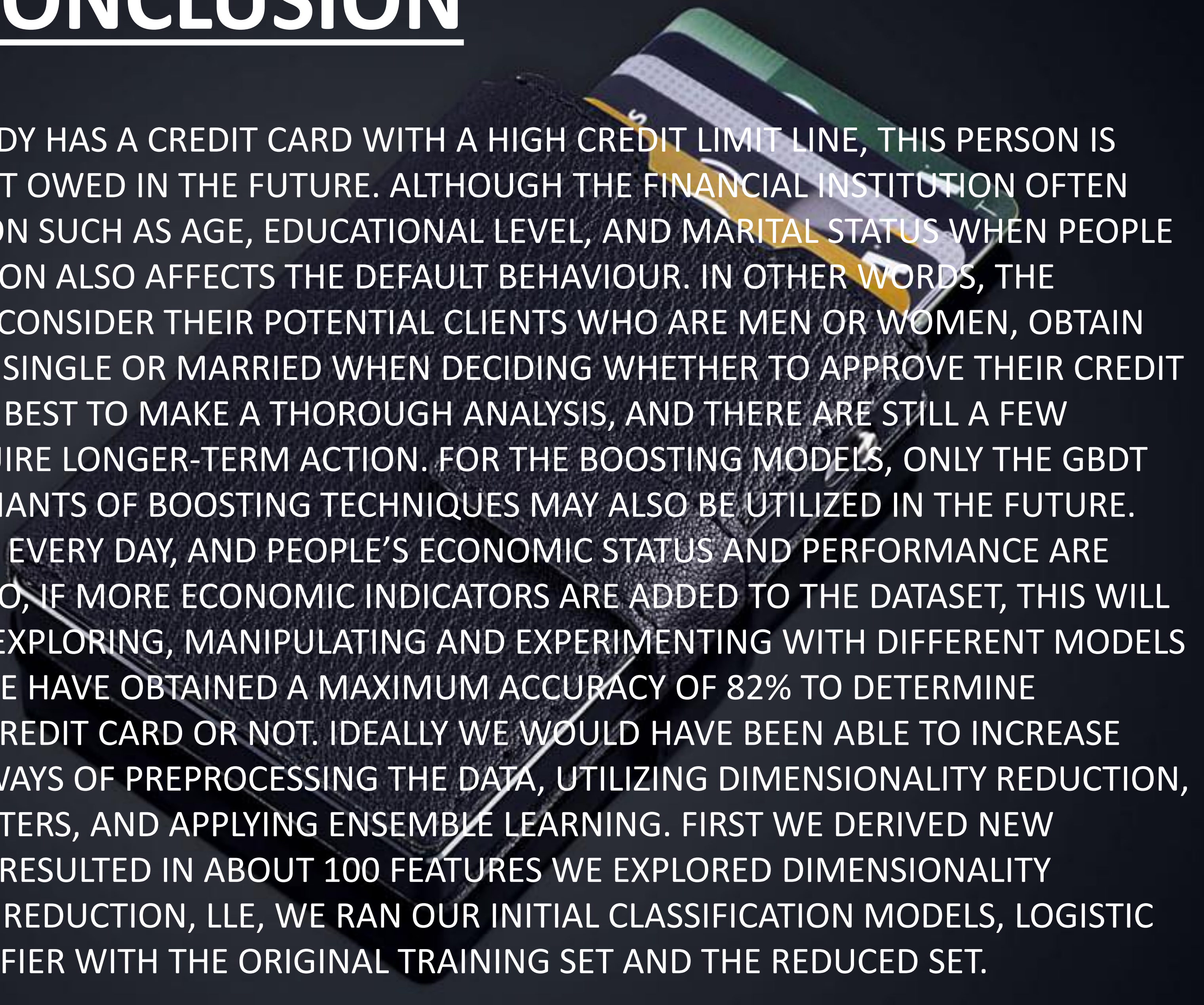
Accuracy of Voting Classifier: 0.8207

CONCLUSION



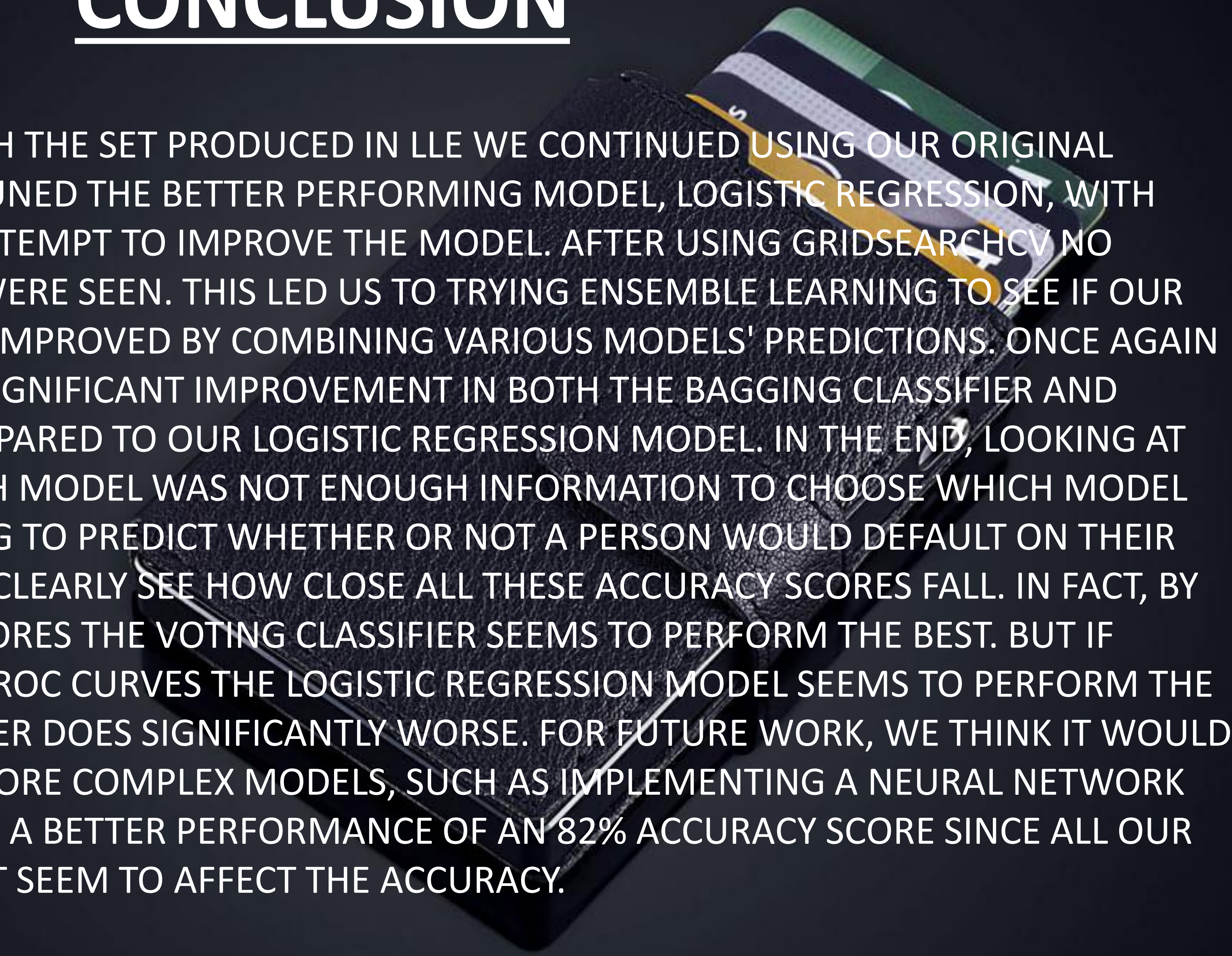
THE OBJECTIVE OF THIS PROJECT IS TO TRAIN VARIOUS SUPERVISED LEARNING ALGORITHMS TO PREDICT THE CLIENT'S BEHAVIOUR IN PAYING OFF THE CREDIT CARD BALANCE. IN CLASSIFICATION PROBLEMS, AN IMBALANCED DATASET IS ALSO CRUCIAL TO ENHANCE THE PERFORMANCE OF THE MODEL, SO DIFFERENT RESAMPLING TECHNIQUES WERE ALSO USED TO BALANCE THE DATASET. WE FIRST INVESTIGATED THE DATASETS BY USING EXPLORATORY DATA ANALYSIS TECHNIQUES, INCLUDING DATA NORMALIZATION. WE STARTED WITH THE LOGISTIC REGRESSION MODEL, THEN COMPARED THE RESULTS WITH TRADITIONAL MACHINE LEARNING-BASED MODELS. THEN K-MEANS SMOTE RESAMPLING METHOD ON TAIWAN CLIENT'S CREDIT DATASET. IN THE END, THE PROPOSED METHOD HAS ALSO BEEN DEPLOYED ON THE WEB TO ASSIST THE DIFFERENT STAKEHOLDERS. THEREFORE, WHEN THE FINANCIAL INSTITUTION CONSIDERS ISSUING THE CLIENT A CREDIT CARD, THE INSTITUTION NEEDS TO CHECK THE PAYMENT HISTORY OF THAT PERSON BECAUSE THE DECISION ON WHETHER TO PAY ON DULY OR OWE THE BILL ON A SPECIFIC MONTH USUALLY RELATES TO THE PREVIOUS PAYMENT HISTORY. FOR INSTANCE, IF A PERSON OWES NUMEROUS BILLS ALREADY, HE OR SHE IS LIKELY TO DELAY THE PAYMENT OF THE CURRENT MONTH UNLESS THIS PERSON GETS A WINDFALL SO THAT THE TOTAL ARREARS CAN BE PAID OFF. BESIDES THE PAYMENT HISTORY, IT IS ALSO IMPERATIVE TO LOOK AT THE APPLICANTS' CREDIT LIMIT OF THEIR CURRENT CREDIT CARDS. THIS IS A RESULT OF A VIRTUOUS CIRCLE: PEOPLE WHO PAY ON DULY TEND TO HAVE BETTER CREDIT SCORES, SO THE BANKS PREFER TO INCREASE THESE PEOPLE'S CREDIT LINES BY TAKING LESS RISK.

CONCLUSION



AS A RESULT, IF A POTENTIAL CLIENT ALREADY HAS A CREDIT CARD WITH A HIGH CREDIT LIMIT LINE, THIS PERSON IS UNLIKELY TO FAIL TO PAY THE FULL AMOUNT OWED IN THE FUTURE. ALTHOUGH THE FINANCIAL INSTITUTION OFTEN COLLECTS CLIENTS’ PERSONAL INFORMATION SUCH AS AGE, EDUCATIONAL LEVEL, AND MARITAL STATUS WHEN PEOPLE APPLY FOR CREDIT CARDS, THIS INFORMATION ALSO AFFECTS THE DEFAULT BEHAVIOUR. IN OTHER WORDS, THE FINANCIAL INSTITUTION SHOULD EQUALLY CONSIDER THEIR POTENTIAL CLIENTS WHO ARE MEN OR WOMEN, OBTAIN BACHELOR DEGREES OR MASTER DEGREES, SINGLE OR MARRIED WHEN DECIDING WHETHER TO APPROVE THEIR CREDIT CARD/LOAN APPLICATIONS. WE TRIED OUR BEST TO MAKE A THOROUGH ANALYSIS, AND THERE ARE STILL A FEW POSSIBLE IMPROVEMENTS THAT MAY REQUIRE LONGER-TERM ACTION. FOR THE BOOSTING MODELS, ONLY THE GBDT METHOD WAS TRAINED, BUT VARIOUS VARIANTS OF BOOSTING TECHNIQUES MAY ALSO BE UTILIZED IN THE FUTURE. THE FINANCIAL MARKET CHANGES RAPIDLY EVERY DAY, AND PEOPLE’S ECONOMIC STATUS AND PERFORMANCE ARE AFFECTED BY THE MARKET ALL THE TIME. SO, IF MORE ECONOMIC INDICATORS ARE ADDED TO THE DATASET, THIS WILL LEAD TO A MORE GENERIC MODEL. AFTER EXPLORING, MANIPULATING AND EXPERIMENTING WITH DIFFERENT MODELS ON THE CREDIT CARD DEFAULT DATA SET WE HAVE OBTAINED A MAXIMUM ACCURACY OF 82% TO DETERMINE WHETHER A PERSON DEFAULTS ON THEIR CREDIT CARD OR NOT. IDEALLY WE WOULD HAVE BEEN ABLE TO INCREASE THIS ACCURACY BY TRYING OUT VARIOUS WAYS OF PREPROCESSING THE DATA, UTILIZING DIMENSIONALITY REDUCTION, FINE-TUNING THE MODELS' HYPERPARAMETERS, AND APPLYING ENSEMBLE LEARNING. FIRST WE DERIVED NEW FEATURES FROM THE DATA SET. SINCE THIS RESULTED IN ABOUT 100 FEATURES WE EXPLORED DIMENSIONALITY REDUCTION. USING THE SECOND FORM OF REDUCTION, LLE, WE RAN OUR INITIAL CLASSIFICATION MODELS, LOGISTIC REGRESSION AND SUPPORT VECTOR CLASSIFIER WITH THE ORIGINAL TRAINING SET AND THE REDUCED SET.

CONCLUSION



SEEING NO IMPROVEMENT WITH THE SET PRODUCED IN LLE WE CONTINUED USING OUR ORIGINAL TRAINING SET. NEXT WE FINE-TUNED THE BETTER PERFORMING MODEL, LOGISTIC REGRESSION, WITH GRIDSEARCHCV AS ANOTHER ATTEMPT TO IMPROVE THE MODEL. AFTER USING GRIDSEARCHCV NO SIGNIFICANT IMPROVEMENTS WERE SEEN. THIS LED US TO TRYING ENSEMBLE LEARNING TO SEE IF OUR OVERALL ACCURACY COULD BE IMPROVED BY COMBINING VARIOUS MODELS' PREDICTIONS. ONCE AGAIN THERE DID NOT SEEM TO BE A SIGNIFICANT IMPROVEMENT IN BOTH THE BAGGING CLASSIFIER AND VOTING CLASSIFIER WHEN COMPARED TO OUR LOGISTIC REGRESSION MODEL. IN THE END, LOOKING AT THE ACCURACY SCORES OF EACH MODEL WAS NOT ENOUGH INFORMATION TO CHOOSE WHICH MODEL PERFORMED BEST WHEN TRYING TO PREDICT WHETHER OR NOT A PERSON WOULD DEFAULT ON THEIR CREDIT CARD. ABOVE ONE CAN CLEARLY SEE HOW CLOSE ALL THESE ACCURACY SCORES FALL. IN FACT, BY JUST SEEING THE ACCURACY SCORES THE VOTING CLASSIFIER SEEMS TO PERFORM THE BEST. BUT IF LOOKED AT PERFORMANCE VIA ROC CURVES THE LOGISTIC REGRESSION MODEL SEEMS TO PERFORM THE BEST AND THE VOTING CLASSIFIER DOES SIGNIFICANTLY WORSE. FOR FUTURE WORK, WE THINK IT WOULD BE INTERESTING TO DEVELOP MORE COMPLEX MODELS, SUCH AS IMPLEMENTING A NEURAL NETWORK AND SEEING IF THERE COULD BE A BETTER PERFORMANCE OF AN 82% ACCURACY SCORE SINCE ALL OUR PREVIOUS TECHNIQUES DID NOT SEEM TO AFFECT THE ACCURACY.

THANK
YOU