Credit Card Default Detection

Group 11: *Gun Kang(Coding), Lauren Mieczkowski(Report and Presentation), Virinchi Sharma(Coding), Youngjin Kim(Report and Presentation)*

Project Introduction

Credit cards enable consumers to finance purchases with short terms loans with relative ease. With its invention, we are heading towards a "cashless society" in which most transactions are made based on an individual's credibility. Testifying to the prevalence of this type of transaction, as of June 2018, 83% of Americans, age 30-49, had credit cards (Statistia, 2019). However, some people fail to fulfill expected credibility. In the United States, as of the second quarter of 2019, the total credit card debt reached about 0.87 trillion USD (Statistia, 2019). Having debt not only decreases an individual's credit score, a metric that determines other loan opportunities but also, decreases consumption growth which negatively affects the economy (Tufan Ekici & Lucia Dunn, 2010). In the interest of companies who strive to give credit to trustworthy customers and individuals who aim to achieve financial well-being, an inquisition into what determines credit card default is substantiated. In this project, we will investigate the attributes that contribute to credit card default. Based on our models, we can properly classify candidates and determine the importance of various characteristics.

Dataset Description

Our dataset, "Default of Credit Card Clients Data", originates from the UCI machine learning repository based upon a study of Taiwan default payments in 2005 and was donated in 2016. There is a total of 30,000 rows and 24 columns. It includes variables such as the total amount of given credit, gender, education level, marriage status, age, 6 months of past payments, 6 months of bill statements, 6 months of previous payments, and lastly, whether or not it was a default payment.

The detailed attribute information is as follows, sourced from the UCI Machine Learning Repository, 2016:

- LIMIT_BAL: Amount of given credit (NT dollar), which includes both individual consumer credit and family (supplementary) credit. For ease of understanding, this is transformed into USD circa 2005 using the federal reserve's published conversion rate.
- SEX: Gender (1 = Male / 2 = Female) This variable is transformed into a dummy variable which has 0 as Female and 1 as Male.

- EDU: Education (1 = Graduate school / 2 = University / 3 = High School / 4 = Others)
  We will be separating this attribute into four different binary dummy variables (educ_grad, educ_univ, educ_hs, educ_other) which represent respective education level. Education other may represent primary school or middle school.
- MARRIAGE: Marital Status (1 = Married / 2 = Single / 3 = Others)
  We will be separating this attribute into three different binary dummy variables (marriage_married, marriage_single, and marriage_others) which represent respective marital status. Marriage others may represent divorced or dating.
- AGE: Age in years.
- PAY_0 to PAY_6: Monthly repayment status from September, 2005 to April, 2005, where PAY_0 represents the repayment status in September, 2005 and Pay_6 represents the repayment status in April, 2005. (-1 = Pay duly / 1 = Payment delay for one month / 2 = Payment delay for two months / … / 9 = Payment delay for nine months or above.) As the number increases, there is more delay in payment.
- BILL_AMT1 to BILL_AMT6: Amount of bill statement (NT dollar) from September, 2005 to April, 2005. For ease of understanding, this is transformed into USD circa 2005 using the federal reserve's published conversion rate.
- PAY_AMT1 to PAY_AMT6: A of previous payment (NT dollar) from September, 2005 to April, 2005. For ease of understanding, this is transformed into USD circa 2005 using the federal reserve's published conversion rate.
- DEFAULT.PAYMENT.NEXT.MONTH: Default Status (1= Default / 0= No Default)
  We will refer this variable as "default". Default means failure to pay upon the agreed terms.

Methodologies

In the data preprocessing stage, we will be looking for missing values and take corresponding measures to replace or remove the missing values. Moreover, we will be separating the education variable into four different binary variables (educ_grad, educ_univ, educ_hs, educ_other) and the marraige variable into three binary variables (marriage_married, marriage_single, and marriage_others). In addition, sex is also changed to a binary variable. Using these dummy variables will enable us to see how each categorical variable affects our dependent variable "default" in comparison to a reference category.

After the data preprocessing, we will be able to explore the dataset by visualizing barplots for "default" with regards to variables like EDU and AGE. Histograms regarding the continuous variables will provide insight into their distribution, to aid in proper modeling. These visualizations will assist our understanding of the dataset by providing an overlook.

Next, allowing for unbiased assessment of the final model, we split the data into a training dataset and test dataset of 75% and 25% of the observations. This procedure is done using stratify which will maintain the target proportion of the original dataset to ward against further imbalance in the dependent variable. Based on the training dataset, different classification techniques, such as logistic regression, decision tree, random forest, ensemble, and neural network, will be utilized to build supervised predictive models. Considering an imbalance of target data, cross-validation of 5 folds will be used to train a more reliable model. This means that our original training data is further divided into 5 folds for training and validation. These folds are then distributed to 5 different sections, each with different combinations of 4 parts training and 1 validation. Each section builds the model using the training data and applies to towards the validation data. The final performance measures of the model are then calculated as an average of these validated scores. Moreover, we use the gridsearchcv function to tune hyperparameters and obtain a model that best fits our training dataset. During this process, for each set of specifications the cv score is determined and the model associated with the highest f1 score is chosen.

To elaborate, a logistic regression predicts the probability of an event by fitting data to a logit function that uses maximum likelihood estimation. By looking at multicollinearity and statistical significance, an attribute transformation and selection process takes place to maintain the model's integrity. The resulting coefficients of the final model are then transformed into odds ratios to be interpreted. By contrast, decision trees split attributes through nodes and branch out to leaves in order to purely divide the target variables. With this technique, the mathematical measure of disorder called entropy will be applied. By adopting this calculation as the split selection criterion, the tree will then be developed based on highest information gain, the reduction of impurity. The maximum and minimum of different criterion such as the nodes, splits, and depth can be determined. Following suit, a random forest model is built using many random unpruned decision trees and aggregating them together. Notably, each tree uses a different training sample and within, each node has a randomized subset of the attributes to base the split upon. By essentially revealing a majority vote within the trees, prediction accuracy and efficiency are enhanced. Our ensemble model, a bagged decision tree, is similar. Each decision tree version is trained on different training bootstrap samples, which use replacement to make a sample of similar size to the original. By contrast, all attributes are available to split the nodes and then the trees are combined together to create the final model. Finally, a neural network model is developed to consider non-linear functions of the input. A neural network is a contemporary approach to classification based on the brain. It is composed of feed forward layers of varying nodes that pass data transformed with each nodes' corresponding weights. Only when a predetermined threshold value is reached, the nodes then shoot the number forward through the layers. While training the model, these thresholds and weights are first randomized and then tuned to become the key in the mathematical metamorphosis of input data to output. Each iteration of updating the weights during training is called an epoch, while the mathematical

functions that defines the transformation within each node is called an activation function. Lastly, the hidden layers are the layers between input and output. Each of these preceding parameters can be adjusted to attain the optimal neural network.

These models will then be employed to predict the test dataset. After, we will be able to observe the model's performance. Keeping in mind that our dependent variable is imbalanced in nature, with only 22% recorded as default, we cannot use only accuracy rate to evaluate model performance. Large true negative counts utilized within this metric can cause misleading accuracy scores. It is then necessary to evaluate the model through a confusion matrix, which composes a table of true positive, true negative, false positive, and false negative counts. Further, precision, a score of the accuracy for those deemed positive and recall, a score of how well the class of interest was detected, will be shown. These measures are key in evaluating target variables with imbalance. Lastly, the ROC curve and its AUC will simplify the comparison among the performance of five different models that we have used. The receiving operation curve is a measure of proper classification which graphs the model's tradeoff between sensitivity, percent of true positives correctly identified, and 1-specificity, false positive rate in which true negatives are incorrectly identified. So, the area under the curve captures how well the model is predicting overall.

As a result, the model with the highest AUC will be selected as the best predictive model that could be utilized in the future when a new dataset is given, with special attention given to recall scores.

Data Preprocessing and Exploration

Figure 1. Missing Values

Out[11]:

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	female	educ_other	educ_grad	educ_univ	educ_hs	educ_l
0	1	20000.0	2	2	1	24	2	2	-1	-1	...	1	0	0	1	0	
1	2	120000.0	2	2	2	26	-1	2	0	0	...	1	0	0	1	0	
2	3	90000.0	2	2	2	34	0	0	0	0	...	1	0	0	1	0	
3	4	50000.0	2	2	1	37	0	0	0	0	...	1	0	0	1	0	
4	5	50000.0	1	2	1	57	-1	0	-1	0	...	0	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
29995	29996	220000.0	1	3	1	39	0	0	0	0	...	0	0	0	0	1	
29996	29997	150000.0	1	3	2	43	-1	-1	-1	-1	...	0	0	0	0	1	
29997	29998	30000.0	1	2	2	37	4	3	2	-1	...	0	0	0	1	0	
29998	29999	80000.0	1	3	1	41	1	-1	0	0	...	0	0	0	0	1	
29999	30000	50000.0	1	2	1	46	0	0	0	0	...	0	0	0	1	0	

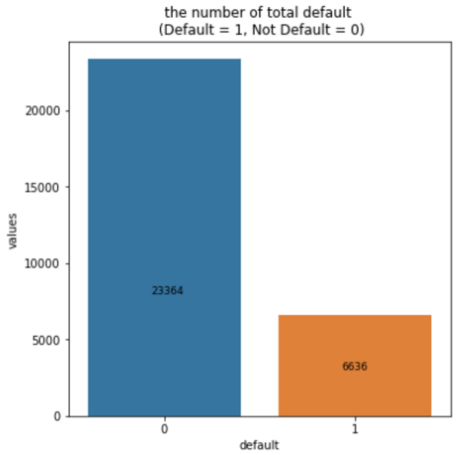
30000 rows × 36 columns

Check for null values

```
In [3]: data.isnull().values.any()
Out[3]: False
```

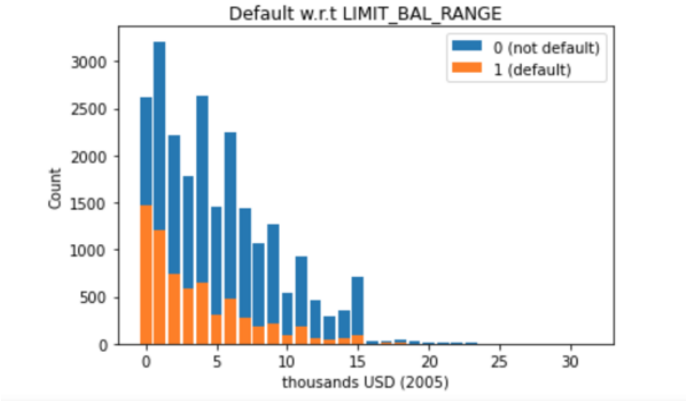
Firstly, after checking for missing observations, it is evident that the UCI Credit Card dataset contains 30,000 rows of observations with no missing values.

Figure 2. Default Barplot



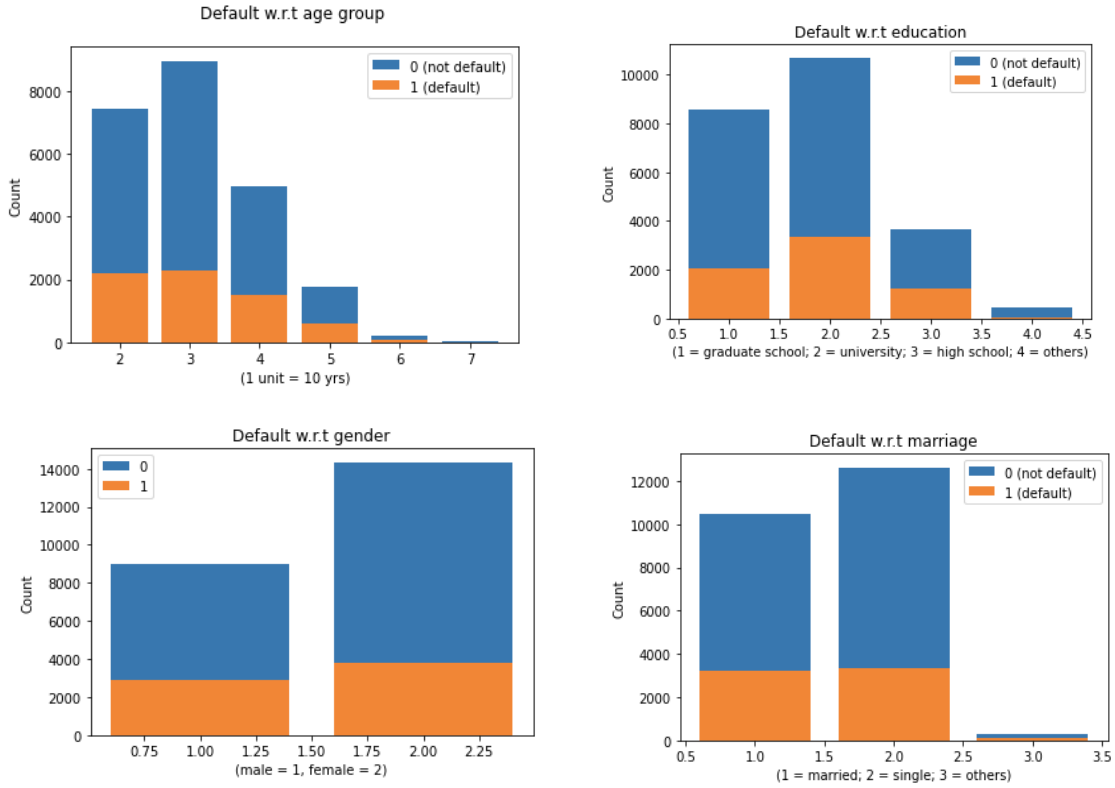
As demonstrated in figure 1, the dependent variable in question has an imbalance of data with 23364 non-defaults (77.9%) in comparison to 6636 observations (22.1%) with default status. This disproportion is recognized through the methods used later in the splitting of data, the building of the models and their evaluation.

Figure 3. Balance Limit and Default Status Barplots



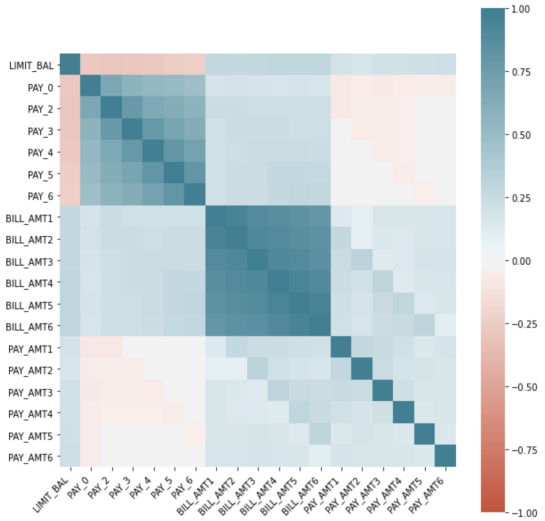
Exploring the relationship between balance limit and the default status, as the limit lowers there seems to be a larger proportion of defaults. Perhaps credit card companies tend to lower balance limits for customers with low credit scores, who's bad credit reputation is then affirmed again through defaulting. The majority of consumers also have lower balance limits.

Figure 4. Categorical Variables and Default Status Barplots



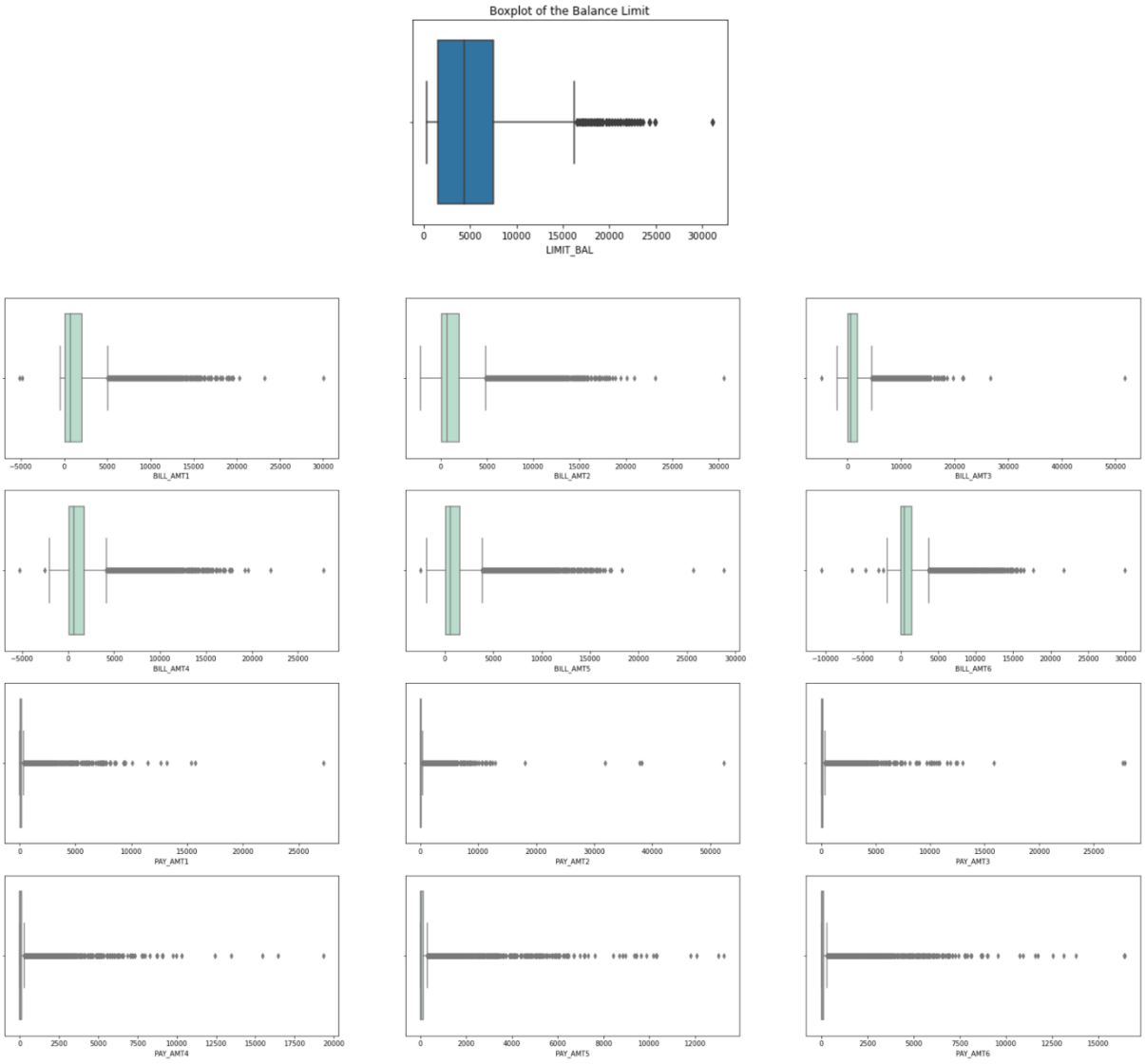
Examining the categorical variables and default status, there are no obvious disparities. By age group, the default proportions seem to be relatively constant with more young people having credit cards. Next, there is a large amount of university graduates with relatively even proportions. In regards to gender, women seem to default more often than men. Lastly, married consumers seem to default less than singles.

Figure 5. Correlation Heat Map



In figure 5, a correlation heat map displays the strength of the relationships between the continuous variables. Notably, bill amounts and pay amounts are strongly correlated within their variable group. With a negative relationship, payment indicators like pay_0 seem to have a medium correlation with limit balance. All other correlations shown are weak.

Figure 6. Boxplots of Continuous Variables



In order to show the distributions of the continuous variables of balance limit, bill amounts and payment amounts, boxplots are charted in USD. As made evident above, there is great disparity in these distributions with a great number of right-hand outliers. These distributions are skewed.

Results

MODEL I. LOGISTIC REGRESSION

We will be utilizing a logistic regression to estimate the relationships between the dependent variable of default, and various independent variables to find out which have the strongest impact. Because we have categorical variables, the dummy variables of education other and marriage other have been excluded from the model as a reference point.

The regression model to be estimated is as follows:

$$Default = \alpha + B_1 Limit\_Bal + B_2 Age + B_3 Pay_0 + B_4 Pay_2 + B_5 Pay_3 + B_6 Pay_4 + B_7 Pay_5 + B_8 Pay_6 + B_9 Bil\_Amt1 + B_{10} Bil\_Amt2 + B_{11} Bil\_Amt3 + B_{12} Bil\_Amt4 + B_{13} Bil\_Amt5 + B_{14} Bil\_Amt6 + B_{15} Pay\_Amt1 + B_{16} Pay\_Amt2 + B_{17} Pay\_Amt3 + B_{18} Pay\_Amt4 + B_{19} Pay\_Amt5 + B_{20} Pay\_Amt6 + B_{21} Male + B_{22} Single + B_{23} Married + B_{24} Educ\_Grad + B_{25} Educ\_Univ + B_{26} Educ\_Hs + \mu$$

Because the continuous variables of limit balance, payment amounts and bill amounts have extreme outliers, these will be standardized so as to not obtain misleading results. The logistic regression was run with and without these standardized variables in order to demonstrate the effect.

Figure 7. Logistic Regression Results and Odds Ratio

Logit Regression Results

Dep. Variable:	default.payment.next.month	No. Observations:	22500
Model:	Logit	Df Residuals:	22474
Method:	MLE	Df Model:	25
Date:	Mon, 30 Nov 2020	Pseudo R-squ.:	0.1192
Time:	23:46:14	Log-Likelihood:	-10472.
converged:	True	LL-Null:	-11890.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
LIMIT_BAL	-2.964e-05	5.91e-06	-5.018	0.000	-4.12e-05	-1.81e-05
AGE	-0.0025	0.002	-1.242	0.214	-0.006	0.001
PAY_0	0.5722	0.020	27.974	0.000	0.532	0.612
PAY_2	0.0853	0.023	3.673	0.000	0.040	0.131
PAY_3	0.0779	0.026	3.001	0.003	0.027	0.129
PAY_4	0.0411	0.029	1.428	0.153	-0.015	0.097
PAY_5	0.0141	0.031	0.454	0.650	-0.047	0.075
PAY_6	0.0155	0.025	0.611	0.541	-0.034	0.065
BILL_AMT1	-0.0001	4.04e-05	-3.647	0.000	-0.000	-6.81e-05
BILL_AMT2	2.244e-05	5.6e-05	0.401	0.689	-8.73e-05	0.000
BILL_AMT3	4.306e-05	5.15e-05	0.837	0.403	-5.78e-05	0.000
BILL_AMT4	1.082e-05	5.07e-05	0.213	0.831	-8.85e-05	0.000
BILL_AMT5	3.047e-05	5.47e-05	0.557	0.577	-7.67e-05	0.000
BILL_AMT6	3.99e-07	4.34e-05	0.009	0.993	-8.47e-05	8.54e-05
PAY_AMT1	-0.0004	8.17e-05	-4.615	0.000	-0.001	-0.000
PAY_AMT2	-0.0004	8.02e-05	-4.408	0.000	-0.001	-0.000
PAY_AMT3	-4.953e-05	6.1e-05	-0.812	0.417	-0.000	7e-05
PAY_AMT4	-0.0001	6.52e-05	-1.716	0.086	-0.000	1.59e-05
PAY_AMT5	-5.186e-05	6.35e-05	-0.817	0.414	-0.000	7.26e-05
PAY_AMT6	-6.737e-05	4.81e-05	-1.401	0.161	-0.000	2.69e-05
male	0.1149	0.035	3.240	0.001	0.045	0.184
single	-0.9807	0.101	-9.704	0.000	-1.179	-0.783
married	-0.7292	0.105	-6.921	0.000	-0.936	-0.523
educ_grad	-0.0117	0.109	-0.107	0.914	-0.226	0.203
educ_univ	-0.1348	0.106	-1.267	0.205	-0.343	0.074
educ_hs	-0.1703	0.114	-1.499	0.134	-0.393	0.052

Odds Ratio

LIMIT_BAL	0.999970
AGE	0.997538
PAY_0	1.772074
PAY_2	1.089073
PAY_3	1.081044
PAY_4	1.041909
PAY_5	1.014149
PAY_6	1.015617
BILL_AMT1	0.999853
BILL_AMT2	1.000022
BILL_AMT3	1.000043
BILL_AMT4	1.000011
BILL_AMT5	1.000030
BILL_AMT6	1.000000
PAY_AMT1	0.999623
PAY_AMT2	0.999646
PAY_AMT3	0.999950
PAY_AMT4	0.999888
PAY_AMT5	0.999948
PAY_AMT6	0.999933
male	1.121732
single	0.375051
married	0.482283
educ_grad	0.988325
educ_univ	0.873897
educ_hs	0.843371

Figure 8. Logistic Regression Results and Odds Ratio with normalized continuous variables

Logit Regression Results

Dep. Variable:	default.payment.next.month			No. Observations:		22500
Model:	Logit			Df Residuals:		22474
Method:	MLE			Df Model:		25
Date:	Mon, 30 Nov 2020			Pseudo R-squ.:		0.1176
Time:	23:46:19			Log-Likelihood:		-10491.
converged:	True			LL-Null:		-11890.
Covariance Type:	nonrobust			LLR p-value:		0.000

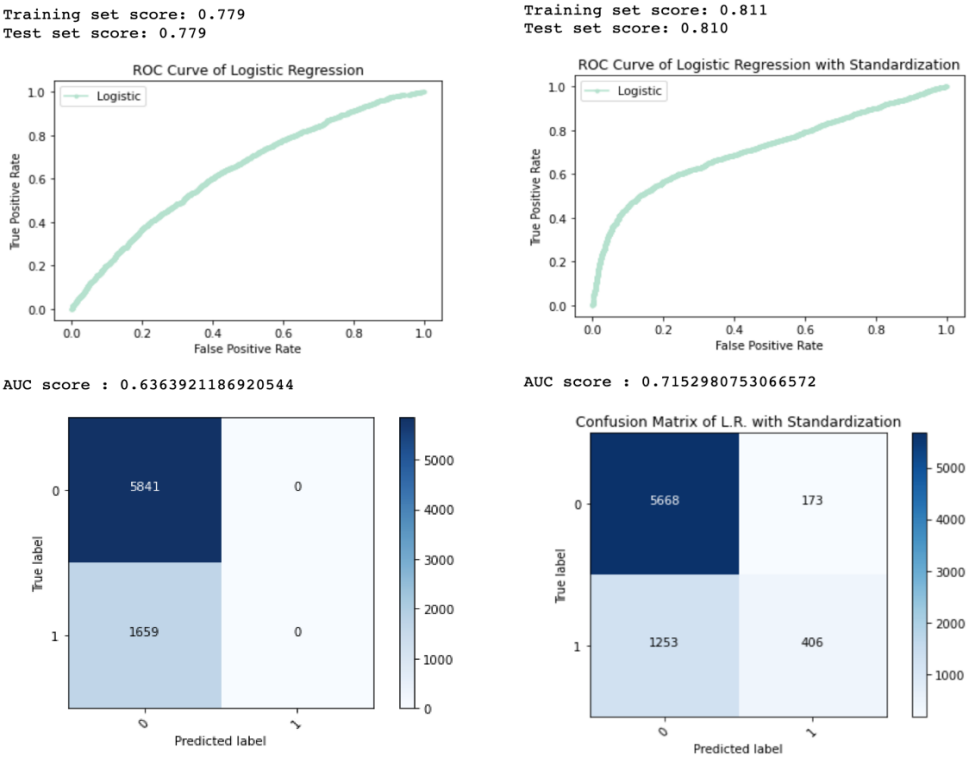
	coef	std err	z	P>|z|	[0.025	0.975]
LIMIT_BAL	-0.0934	0.024	-3.967	0.000	-0.140	-0.047
BILL_AMT1	-0.3268	0.091	-3.586	0.000	-0.505	-0.148
BILL_AMT2	0.0459	0.122	0.375	0.707	-0.194	0.286
BILL_AMT3	0.0878	0.110	0.801	0.423	-0.127	0.303
BILL_AMT4	0.0234	0.100	0.234	0.815	-0.173	0.220
BILL_AMT5	0.0493	0.102	0.482	0.630	-0.151	0.250
BILL_AMT6	0.0051	0.080	0.064	0.949	-0.151	0.161
PAY_AMT1	-0.1848	0.041	-4.499	0.000	-0.265	-0.104
PAY_AMT2	-0.2399	0.056	-4.296	0.000	-0.349	-0.130
PAY_AMT3	-0.0263	0.033	-0.802	0.423	-0.091	0.038
PAY_AMT4	-0.0514	0.031	-1.644	0.100	-0.113	0.010
PAY_AMT5	-0.0252	0.030	-0.847	0.397	-0.084	0.033
PAY_AMT6	-0.0367	0.026	-1.397	0.163	-0.088	0.015
AGE	-0.0045	0.002	-2.299	0.021	-0.008	-0.001
PAY_0	0.5738	0.020	28.055	0.000	0.534	0.614
PAY_2	0.0879	0.023	3.784	0.000	0.042	0.133
PAY_3	0.0793	0.026	3.055	0.002	0.028	0.130
PAY_4	0.0410	0.029	1.427	0.153	-0.015	0.097
PAY_5	0.0155	0.031	0.500	0.617	-0.045	0.076
PAY_6	0.0165	0.025	0.653	0.514	-0.033	0.066
male	0.1160	0.035	3.273	0.001	0.047	0.185
single	-1.1585	0.098	-11.774	0.000	-1.351	-0.966
married	-0.8966	0.103	-8.734	0.000	-1.098	-0.695
educ_grad	-0.1843	0.106	-1.733	0.083	-0.393	0.024
educ_univ	-0.2986	0.104	-2.864	0.004	-0.503	-0.094
educ_hs	-0.3242	0.112	-2.894	0.004	-0.544	-0.105

Odds Ratio

LIMIT_BAL	0.910801
BILL_AMT1	0.721256
BILL_AMT2	1.046970
BILL_AMT3	1.091789
BILL_AMT4	1.023716
BILL_AMT5	1.050570
BILL_AMT6	1.005098
PAY_AMT1	0.831276
PAY_AMT2	0.786675
PAY_AMT3	0.974008
PAY_AMT4	0.949893
PAY_AMT5	0.975102
PAY_AMT6	0.964011
AGE	0.995490
PAY_0	1.774976
PAY_2	1.091838
PAY_3	1.082518
PAY_4	1.041869
PAY_5	1.015590
PAY_6	1.016684
male	1.122941
single	0.313971
married	0.407964
educ_grad	0.831726
educ_univ	0.741863
educ_hs	0.723132

As shown above, standardizing the continuous variables actually made education categories statistically significant. In addition, the coefficients for these continuous variables were increased to a level that is easier to interpret. Lastly, since the odds ratios are Euler's number to the power of the coefficients, their values have changed as well.

Figure 9. Logistic Regression ROC and AUC Comparison



The performance of the logistic model changes significantly. After standardization, the AUC changes from 0.64 to 0.72, indicating an increase in classification performance. This advancement is evident in the shape of the ROC that has moved outwards. Furthermore, before standardization, the model is not predicting any defaults with 0 true positives and 0 true negatives. This is substantially inferior to the model with standardization that indeed predicts 406 defaults correctly.

Figure 10. VIF Scores for Multicollinearity

	feature	VIF
0	LIMIT_BAL	1.579845
1	AGE	1.385131
2	PAY_0	1.919991
3	PAY_2	3.175422
4	PAY_3	3.658305
5	PAY_4	4.287850
6	PAY_5	4.725465
7	PAY_6	3.256654
8	BILL_AMT1	14.036195
9	BILL_AMT2	25.866616
10	BILL_AMT3	21.782219
11	BILL_AMT4	20.349079
12	BILL_AMT5	25.005017
13	BILL_AMT6	15.043005
14	PAY_AMT1	1.708412
15	PAY_AMT2	2.237360
16	PAY_AMT3	1.758122
17	PAY_AMT4	1.648807
18	PAY_AMT5	1.688175
19	PAY_AMT6	1.170035
20	male	1.025608
21	single	20.894018
22	married	20.463188
23	educ_grad	15.437864
24	educ_univ	16.625231
25	educ_hs	9.749443

Next, in order to abide by regression model assumptions, VIF scores are produced to test for multicollinearity. This phenomenon occurs when the independent variables are highly correlated with each other, as shown in the earlier heat map. Multicollinearity is natural in dummy variables but should be avoided otherwise because it affects the precision of the coefficients and can result in inaccurate statistical significance. As shown above, all continuous variables are beneath our standard score of 10 besides billing amounts. Therefore, only the first billing amount will be retained.

The new regression model to be estimated is as follows:

$$Default = \alpha + B_1 Limit\_Bal + B_2 Age + B_3 Pay_0 + B_4 Pay_2 + B_5 Pay_3 + B_6 Pay_4 + B_7 Pay_5 + B_8 Pay_6 + B_9 Bil\_Amt1 + B_{10} Pay\_Amt1 + B_{11} Pay\_Amt2 + B_{12} Pay\_Amt3 + B_{13} Pay\_Amt4 + B_{14} Pay\_Amt5 + B_{15} Pay\_Amt6 + B_{16} Male + B_{17} Single + B_{18} Married + B_{19} Educ\_Grad + B_{20} Educ\_Univ + B_{21} Educ\_Hs + \mu$$

Figure 11. Final Logistic Model Regression Results

Logit Regression Results

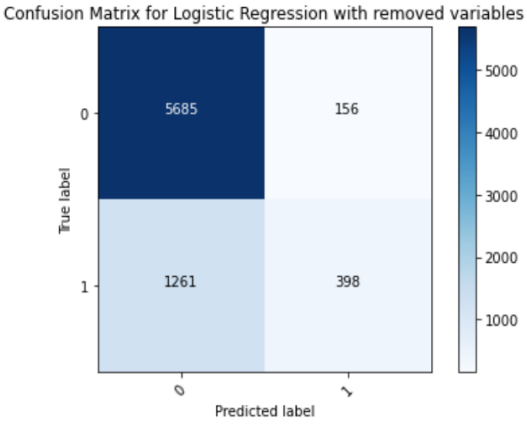
Dep. Variable:	default.payment.next.month	No. Observations:	22500
Model:	Logit	Df Residuals:	22479
Method:	MLE	Df Model:	20
Date:	Mon, 30 Nov 2020	Pseudo R-squ.:	0.1172
Time:	23:46:22	Log-Likelihood:	-10496.
converged:	True	LL-Null:	-11890.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]	Odds Ratio
LIMIT_BAL	-0.0887	0.023	-3.794	0.000	-0.134	-0.043	0.915150
BILL_AMT1	-0.1435	0.023	-6.252	0.000	-0.188	-0.099	0.866318
PAY_AMT1	-0.1576	0.036	-4.331	0.000	-0.229	-0.086	0.854161
PAY_AMT2	-0.2055	0.049	-4.189	0.000	-0.302	-0.109	0.814201
PAY_AMT3	-0.0209	0.027	-0.760	0.447	-0.075	0.033	0.979321
PAY_AMT4	-0.0452	0.027	-1.660	0.097	-0.099	0.008	0.955772
PAY_AMT5	-0.0285	0.024	-1.162	0.245	-0.076	0.020	0.971950
PAY_AMT6	-0.0404	0.026	-1.575	0.115	-0.091	0.010	0.960371
AGE	-0.0045	0.002	-2.270	0.023	-0.008	-0.001	0.995550
PAY_0	0.5777	0.020	28.264	0.000	0.538	0.618	1.781869
PAY_2	0.0839	0.023	3.618	0.000	0.038	0.129	1.087565
PAY_3	0.0825	0.026	3.187	0.001	0.032	0.133	1.085999
PAY_4	0.0437	0.029	1.523	0.128	-0.013	0.100	1.044629
PAY_5	0.0206	0.031	0.670	0.503	-0.040	0.081	1.020842
PAY_6	0.0236	0.025	0.942	0.346	-0.025	0.073	1.023837
male	0.1126	0.035	3.180	0.001	0.043	0.182	1.119176
single	-1.1545	0.098	-11.755	0.000	-1.347	-0.962	0.315201
married	-0.8921	0.103	-8.703	0.000	-1.093	-0.691	0.409781
educ_grad	-0.1804	0.106	-1.700	0.089	-0.388	0.028	0.834921
educ_univ	-0.2970	0.104	-2.854	0.004	-0.501	-0.093	0.743027
educ_hs	-0.3246	0.112	-2.903	0.004	-0.544	-0.105	0.722796

For notable variables, the odds ratios are interpreted to describe the effect of the variables on the probability of defaulting. Only the payment indicators and gender variables have a positive effect on defaulting. Being male in comparison to being female is associated with an increase in odds of defaulting by a factor 1.12 or a 11.92% increase. In comparison to other payment indicators, Pay_0 has the greatest effect: a 1 unit increase is associated with an increase in odds of defaulting by a factor 1.78 or a 78.19% increase. With the second greatest magnitude of all variables, being single in comparison to a status of other is associated with a decrease in odds of defaulting by a factor 0.315 or a 68.48% decrease. Having the greatest effect among education categories, Educ_Hs in comparison to other education is associated with a decrease in odds of defaulting by a factor of 0.72 or a 27.73% decrease. With low magnitude, a 1 unit increase in age is associated with a decrease in odds of defaulting by a factor of 0.99 or a 0.45% decrease. A dollar increase in limit balance is associated with a decrease in odds of defaulting by a factor of 0.92 or a 8.49% decrease. Furthermore, a dollar increase in the first billing amount is associated with a decrease in odds of defaulting by a factor of 0.87 or a 13.37% decrease. With the greatest magnitude among payment amounts, Pay_Amt2 is associated with a decrease in odds of defaulting by a factor of 0.81 or a 18.58% decrease. Each of these odds ratios are made holding all other variables constant.

These odds ratios are important as our final logistic model has statistical significance at the 10% level for all variables except for Pay_Amt3, Pay_Amt5, Pay_Amt6, Pay_4, Pay_5, and Pay_6.

Figure 12. Final Logistic Regression Model Performance Evaluators



	precision	recall	f1-score	support
0	0.82	0.97	0.89	5841
1	0.72	0.24	0.36	1659
accuracy			0.81	7500
macro avg	0.77	0.61	0.62	7500
weighted avg	0.80	0.81	0.77	7500

The AUC score of the logistic regression model is now 0.714. Most importantly, the model is predicting 554 defaults out of 1659, with 398 being correct. Correspondingly, the precision score is 72%. By contrast, the recall score is 24% which means of all true defaults, only 24% of them were identified.

MODEL II. DECISION TREE

Employing a different kind of classifier, a decision tree is computed after tuning parameters. The variables are not transformed, and none are excluded. Using entropy as the criterion for building, the grid search with cross validation indicated the optimal parameters resulting in the highest f1 scores are max depth 4 and max leaf nodes 20. F1 score, calculated using precision and recall, was used instead of accuracy because it provides a better sense of performance for imbalanced classes. In the decision tree, there are 16 leaves with a height of 4 and 15 splits. For each node, a corresponding entropy score is displayed. A low entropy score is desired as it indicates a pure division.

Figure 13. Decision Tree

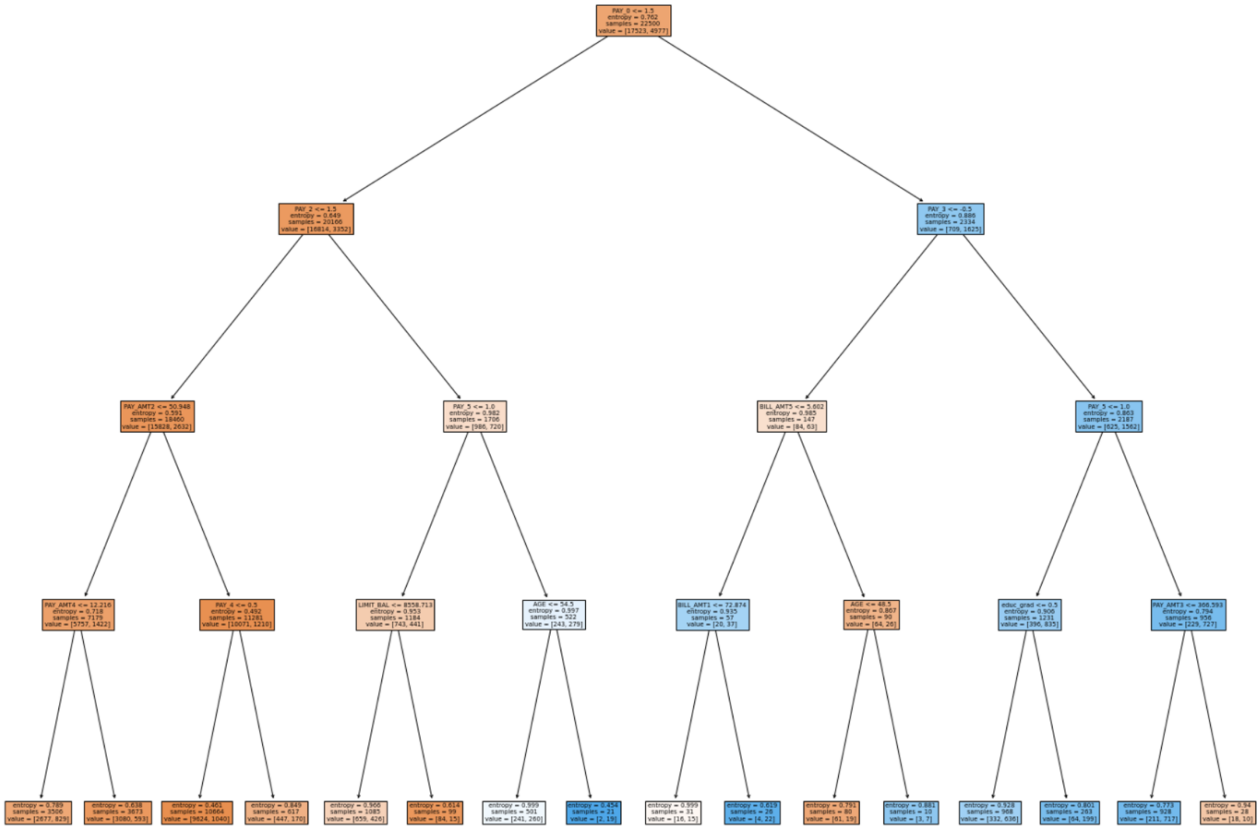
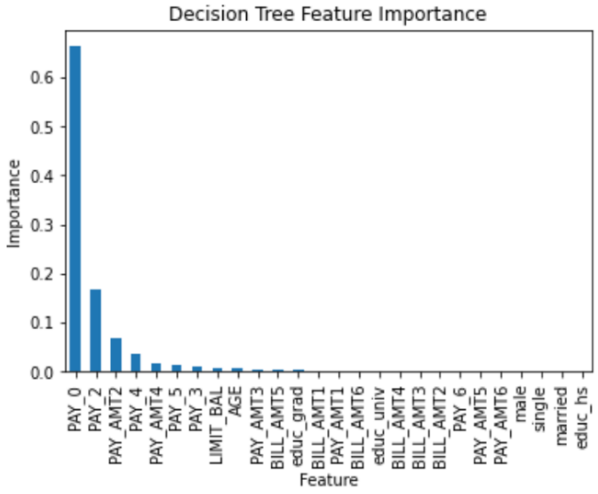


Figure 14. Decision Tree Feature Importance

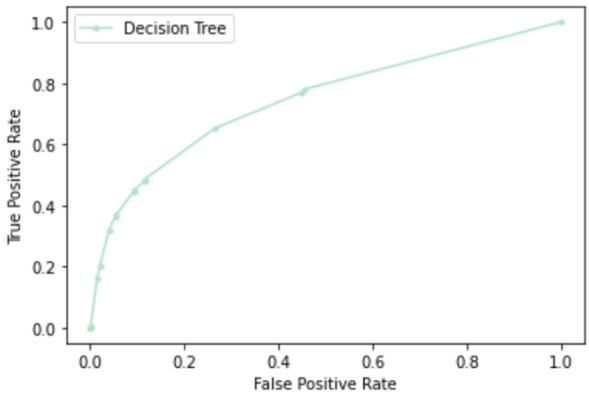


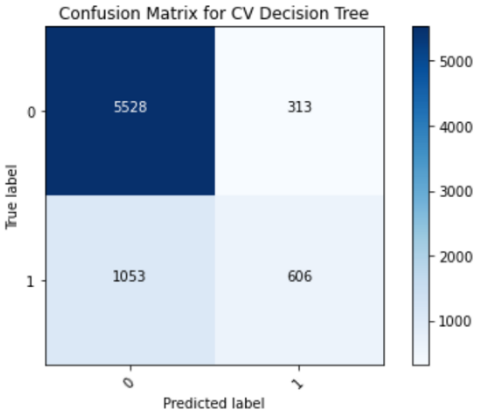
	Importance
PAY_0	0.663403
PAY_2	0.166852
PAY_AMT2	0.069160
PAY_4	0.034707
PAY_AMT4	0.015257
PAY_5	0.013581
PAY_3	0.011593
LIMIT_BAL	0.006124
AGE	0.005370
PAY_AMT3	0.005075
BILL_AMT5	0.004478
educ_grad	0.002339
BILL_AMT1	0.002061
PAY_AMT1	0.000000
BILL_AMT6	0.000000
educ_univ	0.000000
BILL_AMT4	0.000000
BILL_AMT3	0.000000
BILL_AMT2	0.000000
PAY_6	0.000000
PAY_AMT5	0.000000
PAY_AMT6	0.000000
male	0.000000
single	0.000000
married	0.000000
educ_hs	0.000000

As displayed above, gender, relationship status, and bill amounts are not utilized in this decision tree. Of the attributes included, the first month payment delay indicator is the most

important by a large amount. In this context, importance is the computation of information gain weighted by the probability of arriving at the specific node. This measure not only considers the purity difference but also, recognizes the size of observations which were divided. For example, Pay_0 and Pay_2 are some of the first splits, affecting the most observations and having lower entropy at the same time to produce a high importance score. On the other hand, Pay_3 has a split near the beginning, but its importance is not as great as it has fewer samples and high entropy.

Figure 15. Decision Tree Performance Evaluators






```
AUC score : 0.7440432460814352
```

	precision	recall	f1-score	support
0	0.84	0.95	0.89	5841
1	0.66	0.37	0.47	1659
accuracy			0.82	7500
macro avg	0.75	0.66	0.68	7500
weighted avg	0.80	0.82	0.80	7500

The AUC score of the decision tree model is 0.744. Most importantly, the model is predicting 919 defaults out of 1659, with 606 being correct. Correspondingly, the precision score is 66%, which means of the ones predicted to default, 66% of them were correct. By contrast, the recall score is 37% which means of all true defaults, only 37% of them were identified. As the recall score improves, the precision score tends to decrease because of a tradeoff between false negatives and false positives.

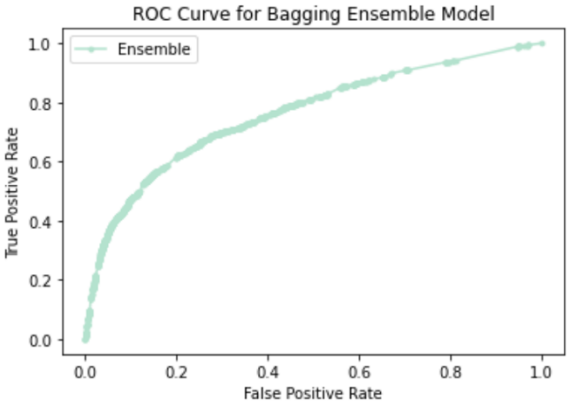
MODEL III. ENSEMBLE

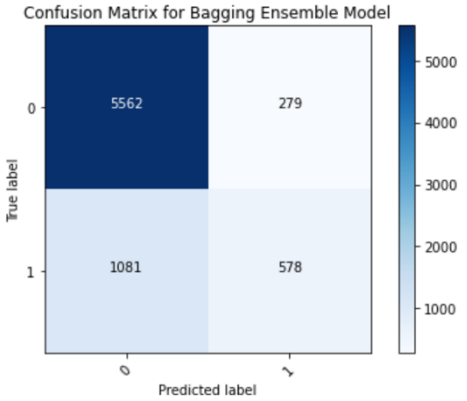
An extension of the decision tree, our ensemble model is a bagged decision tree which combines bootstrap samples and aggregation. Based on f1 scores, the grid search cv function was utilized to select the number of estimators and maximum sample proportion of the bagging

classifier. The bagging classifiers input was the previous optimal decision tree with a max depth of 4 and max leaf nodes of 20. This resulted in the tuned parameters of max samples of 0.2 and max number estimators of 10, meaning that the samples drawn from our training dataset to train each base estimator with replacement are 1/5 the size of the original and that 10 decision trees are aggregated.

Figure 16. Ensemble Model Performance Evaluators

```
AUC score : 0.7629259462557039
```





	precision	recall	f1-score	support
0	0.84	0.95	0.89	5841
1	0.67	0.35	0.46	1659
accuracy			0.82	7500
macro avg	0.76	0.65	0.68	7500
weighted avg	0.80	0.82	0.80	7500

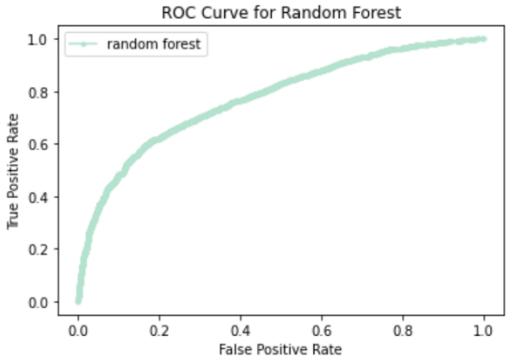
The AUC score of the ensemble model is 0.763. Most importantly, the model is predicting 857 defaults out of 1659, with 578 being correct. Correspondingly, the precision score is 67%, which means of the ones predicted to default, 67% of them were correct. By contrast, the recall score is 35% which means of all true defaults, only 35% of them were identified.

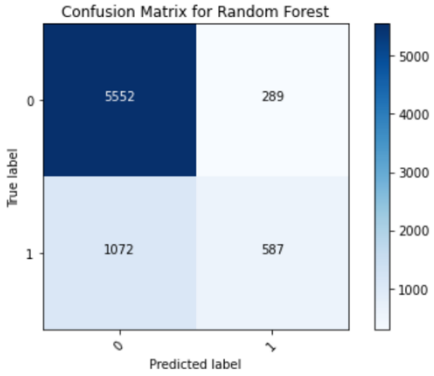
MODEL IV. RANDOM FOREST

Another extension of the decision tree, the random forest model aggregates decision tre es formed with different tuned parameters. After using the grid search and cross validation functi on based on entropy, the optimal maximum features, maximum depth, and number of estimators were 20, 8, and 100 respectively. This translates to 100 aggregated decision trees of height 8 that have a randomized subset of only 20 characteristics to be selected from at each node.

Figure 17. Random Forest Model Performance Evaluators

AUC score : 0.7760554224832277





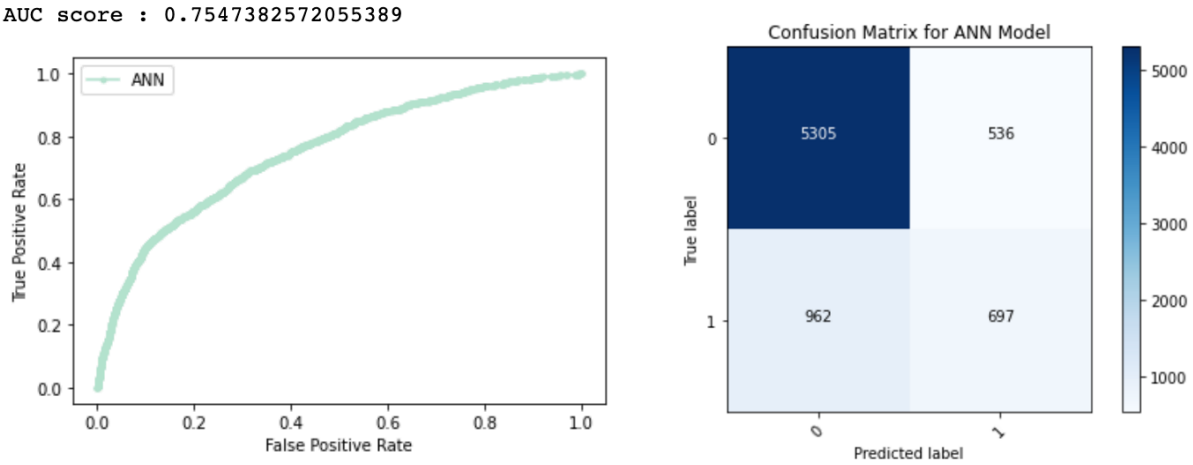
	precision	recall	f1-score	support
0.0	0.84	0.95	0.89	5841
1.0	0.67	0.35	0.46	1659
accuracy			0.82	7500
macro avg	0.75	0.65	0.68	7500
weighted avg	0.80	0.82	0.80	7500

The AUC score of the random forest model is 0.776. Most importantly, the model is predicting 876 defaults out of 1659, with 587 being correct. Correspondingly, the precision score is 67%, which means of the ones predicted to default, 67% of them were correct. By contrast, the recall score is 35% which means of all true defaults, only 35% of them were identified. These results are very similar to the ensemble models scores, which is justified by the similarity in techniques. Random Forest is a more robust version of bagging decision trees as it allows for the sampling of attributes.

MODEL V. NEURAL NETWORK

Utilizing a modern approach to classification, a neural network is created. This model is developed using the grid search cv of 5 folds to determine the optimal layer sizes, node sizes and the activation function with the highest f1 scores. The ideal activation function is the rectified linear activation function, which returns 0 if the input is negative or the original input if positive. Further, the neural network has 5 hidden layers with 20 nodes that has 10000 epochs of updating weights with a constant learning rate. Interestingly, the random forest optimization also resulted in 20 attributes for each node.

Figure 18. Neural Network Model Performance Evaluators



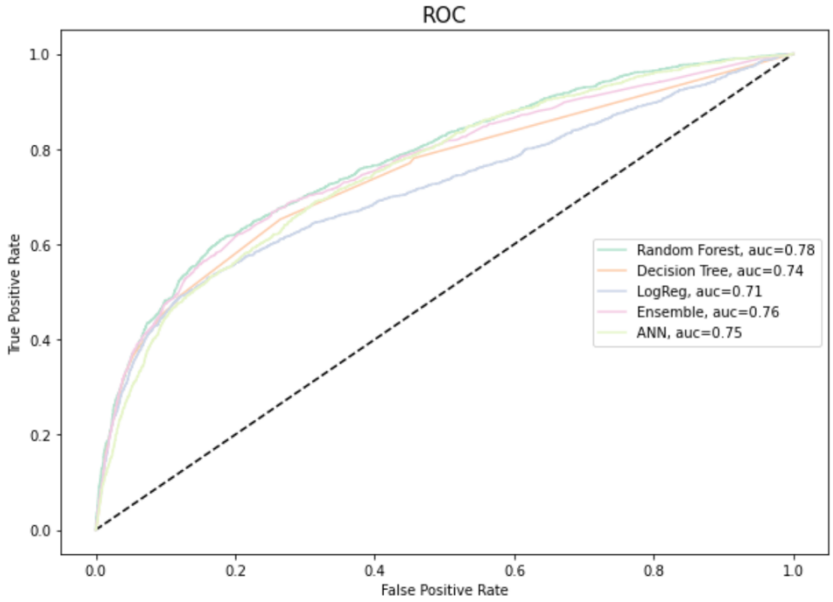
	precision	recall	f1-score	support
0	0.85	0.91	0.88	5841
1	0.57	0.42	0.48	1659
accuracy			0.80	7500
macro avg	0.71	0.66	0.68	7500
weighted avg	0.78	0.80	0.79	7500

The AUC score of the neural network model is 0.755. Most importantly, the model is predicting 1233 defaults out of 1659, with 697 being correct. Correspondingly, the precision score is 57%, which means of the ones predicted to default, 57% of them were correct. By contrast, the recall score is 42% which means of all true defaults, only 42% of them were identified. It is evident that is model classifies substantially more observations as default. In turn, the precision score is decreased but the recall score is increased.

MODEL COMPARISON

In selecting the optimal model, measures that are pertinent to overall classification are valued as well as certain criteria that correspond with the business objectives at hand. Therefore, the AUC score and recall score of each model will be scrutinized. Because defaulting on credit cards is a major financial liability, companies will focus on classifying as many of the true defaults as possible and sacrifice high false positive counts in order to reduce costly false negatives. Once again, accuracy is ignored as true negative scores are relatively invaluable to businesses with high risk and produce misleading results.

Figure 19. ROC Curve Comparison



As can be deciphered above, the random forest model has the highest AUC of all models at 0.78. With more robust features, it is sensible that the random forest classifies better overall than its inferior counterparts of the bagged and normal decision trees. Notably, there is no difference between the recall rates of the ensemble model and random forest.

Having the third highest AUC, the neural network proved to be worthy of deeper investigation. This model had the highest number of true positives among all models, as well as the highest recall rate of 42% in comparison to random forest's recall rate of 35%. The neural network is capturing 7% more of the true defaults correctly, with only a deficit of 0.03 in the AUC.

Furthermore, while standardizing and removing variables improved the logistic model's performance, it still remains the worst classifier of all models. The logistic regression resulted in lowest AUC score and a meager recall rate of only 24%.

Lastly, while the random forest model performs better overall by using the AUC as the standard, the neural network model accomplishes superior classification with recall according to specific business purposes.

Implications

While the neural network and random forest proved to be superior models, their outcomes do not clearly identify the key components of anticipating potential default. Therefore, the decision tree and logistic regression will be utilized to establish the critical factors in predicting whether or not a customer will default on their credit card balance.

Firstly, the logistic regression and decision tree resonate that the first payment indictor is the most crucial. In the decision tree this variable has the greatest magnitude of importance, which upheld in the logistic regression with both statistical significance and the greatest effect on the odds that a customer will default. The magnitude of payment delay at the very beginning of the time period is a great indication of eventual default.

Next, in the decision tree all payment delay indicators except for the 6^{th} month were included. Conversely, the logistic regression revealed no statistical significance for pay_4, pay_5, and pay_6. Of these, the payment indicator for the 4^{th} and 5^{th} month had medium importance on the decision tree. Payment indicator of the second month had the second highest importance while also maintaining a decent effect on the odds of defaulting. Therefore, out of the payment delay indicators, only the first and second month are consistently key in prediction.

According to the logistic regression, the socioeconomic variables such as single, married, high school education, university graduate, and male were all statistically significant with higher effects. Particularly, the relationship status and high school or university education had a greater influence on the odds of defaulting than all variables other than pay_0. By contrast, these variables have low importance in the decision tree, some not even included.

Further, the payment amounts by month have relatively harmonious results as the fifth and sixth month are statistically insignificant with low effects on the odds of defaulting and are not included in the decision tree. Of these, the payment amount of the second month has the highest importance and effect on odds within the category.

Lastly, the limit balance and bill amount 1 are statistically significant but have relatively low importance and low effect on the odds of defaulting. In addition, the inclusion of only the first bill amount in the logistic regression corresponds to the automated selection of just the first bill in the decision tree.

Taking into account these findings, a credit card company who is avoidant of defaulting customers will primarily examine the first and second month payment delay status as well as the second payment amount. There can be some consideration placed towards the 4^{th} payment amount, 3^{rd} payment indicator, limit balance and first billing amount.

Each of these variables relay the kind of risk and financial obligation a customer is comfortable with. Having delays in payment at the beginning of a contract, as well as high charges, demonstrates relative financial irresponsibility that will come into play later when defaulting. Similarly, having a low limit balance can insinuate poor financial well-being which can also be reflected in paying a low amount toward a low bill. But it is not entirely evident with

the data at hand, if a low payment could be towards a high bill. It is also important to recognize revolving utilization, or the proportion of the balance limit used, can also demonstrate monetary liability. In this regard, a person who charges a high proportion of the billing limit may have a financial need that they later cannot fulfill.

In addition, while the demographic indicators are more readily available in assessing potential customers than specific payments, these variables may be hard to rely upon due to conflicting results and may not be discriminated against legally. Therefore, the financial variables should be emphasized in analysis.

In order to fend against defaulting customers, credit card companies first inspect the credit score of customers and then approve credit cards with corresponding terms and conditions, some of which can be modified later. Consumers with good credit will achieve a higher limit balance. The companies should keep in mind that those with lower balance limits have an increase in odds of defaulting, so those who would have been approved with small limits should be under special scrutiny for acceptance. During the beginning of the credit card contract, the behavior of the customer is significant. Credit card companies who notice delays in payment already can increase the interest rate for future balances to dissuade against overspending. Additionally, if the company notices that either the starting billing and payment amounts are low, this may indicate low confidence and ability to pay off the card. The companies could then decrease the balance limit of the customer to prevent spending that cannot be afforded. By paying close attention to these financial variables, customer behavior and preference is revealed to be used in the modification of agreement terms. While earning interest from customers produces income, the risk of costly non-payment ensures that prevention measures against default are utilized.

Sources

Federal Reserve Statistical Release. (2006, January 3). Foreign Exchange Rates (annual) https://www.federalreserve.gov/releases/g5a/20060103/

Statista. (2019, December 10). Topic: Credit cards in the United States. Retrieved September 28, 2020, from https://www.statista.com/topics/1118/credit-cards-in-the-united-states/

Tufan Ekici & Lucia Dunn (2010) Credit card debt and consumption: evidence from household-level data, Applied Economics, 42:4, 455-462, DOI: 10.1080/00036840801964526

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients