CREDIT CARD APPROVAL PREDICTION

UGA Data Science Competition 2021 - Wells Fargo

Undergrad Team: Ayush Kumar, Chloe Phelps, Faisal Hossain, Nicholas Sung

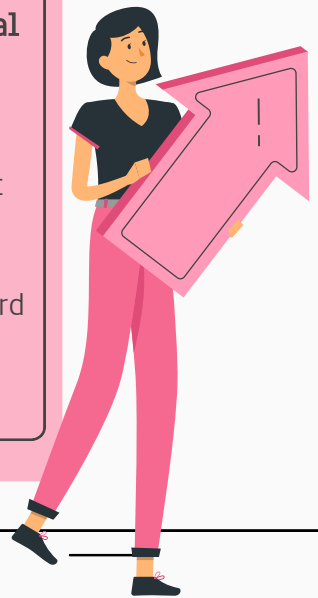
Department of Statistics
Franklin College of Arts and Sciences
UNIVERSITY OF GEORGIA

WELLS FARGO

TABLE OF CONTENTS

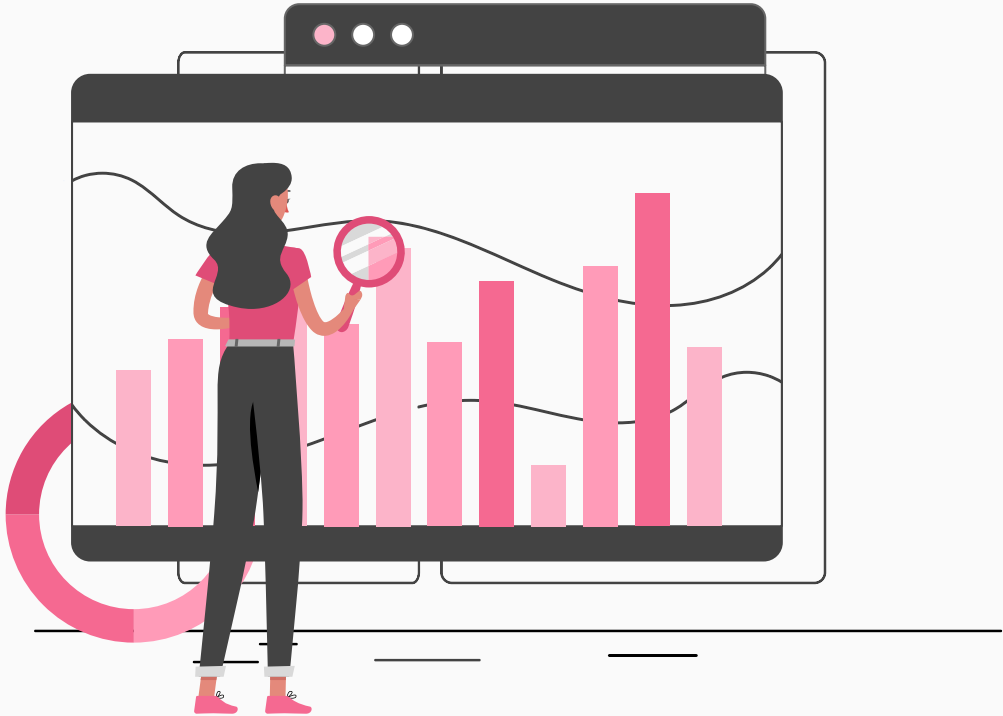
OBJECTIVE

Determine **whether a financial institution should issue** a credit card to an applicant - using personal information and data submitted by credit card applicants in order to predict the probability of future defaults and credit card borrowings.



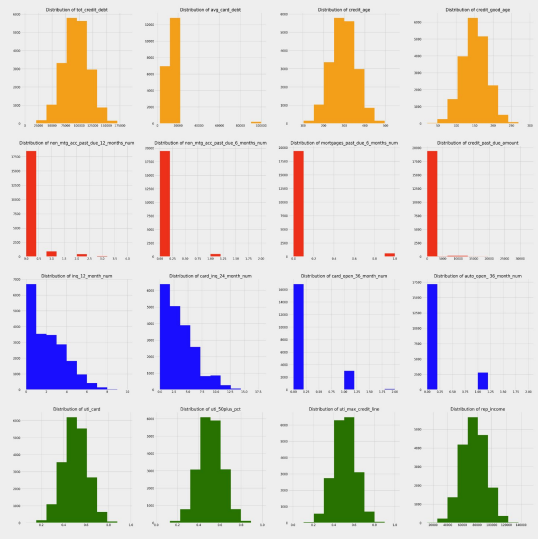
01.
DATASET

Vizulations and Exploratory Analysis
Continuous Variables
Categorical Variables

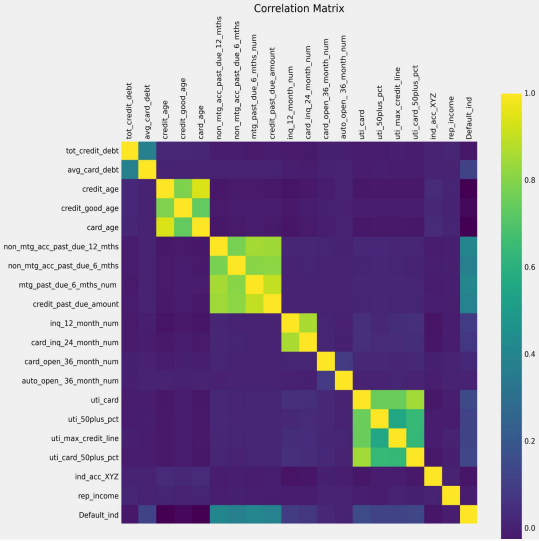


CONTINUOUS VARIABLES

HISTOGRAM



CORRELATION MATRIX



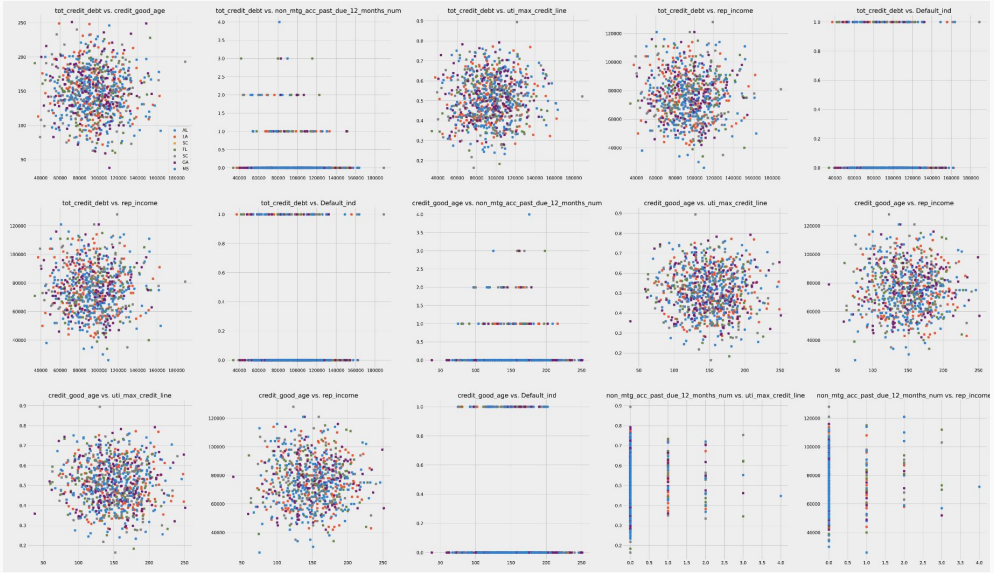
DESCRIPTIVE STATISTICS

	mean	std	min	ma
tot_credit_debt	94563.702530	23546.443862	2367.430000	188890.960000
avg_card_debt	14088.235475	9314.495936	2363.120000	99999.000000
credit_age	296.697000	61.711702	54.000000	545.000000
credit_good_age	149.771750	34.016476	21.000000	296.000000
non_mtg_acc_past_due_12_months_num	0.111350	0.433890	0.000000	4.000000
non_mtg_acc_past_due_6_months_num	0.027400	0.171903	0.000000	2.000000
mortgages_past_due_6_months_num	0.030200	0.171142	0.000000	1.000000
credit_past_due_amount	329.287867	2073.899357	0.000000	32662.980000
inq_12_month_num	1.762700	1.740816	0.000000	10.000000
card_inq_24_month_num	3.409600	2.926697	0.000000	18.000000
card_open_36_month_num	0.163050	0.386099	0.000000	2.000000
auto_open_ 36_month_num	0.141000	0.349607	0.000000	2.000000
uti_card	0.503157	0.109354	0.065120	0.969289
uti_50plus_pct	0.511007	0.113456	0.033749	0.988964
uti_max_credit_line	0.507629	0.108624	0.005174	1.000000
rep_income	75499.511666	16361.955146	12000.000000	150000.000000

- Extreme Min & Max values for Total Credit Debt, Average Credit Debt, and Credit past due amount
- Similar distributions, which may cause multicollinearity
- Multiple points of high collinearity between variables that are similar in nature due to the time that they are recorded or other inherent similarities

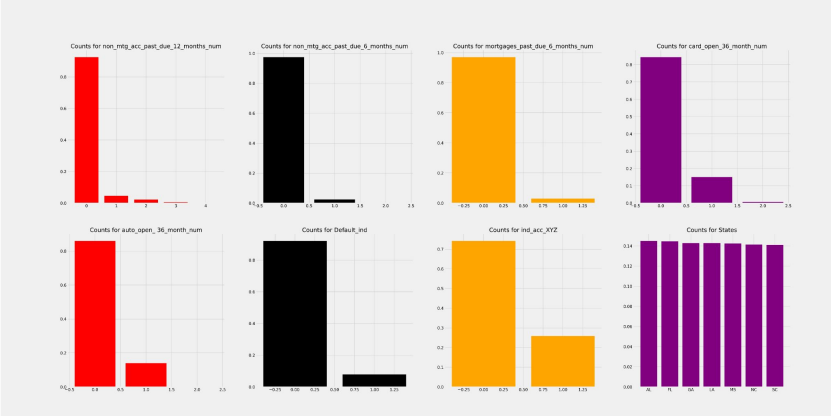
CATEGORICAL VARIABLES

SCATTER PLOTS BY STATE



No strong state interaction effects exists, but some non-linear patterns occur, so the state dummy variables were included in our analysis

COUNTS FOR CATEGORICAL VARIABLES



Counts for these categorical variables exhibit the relative frequencies with which they occur. Dummy variables will be included for all of these variables at multiple levels.

02.
LOGISTIC REGRESSION

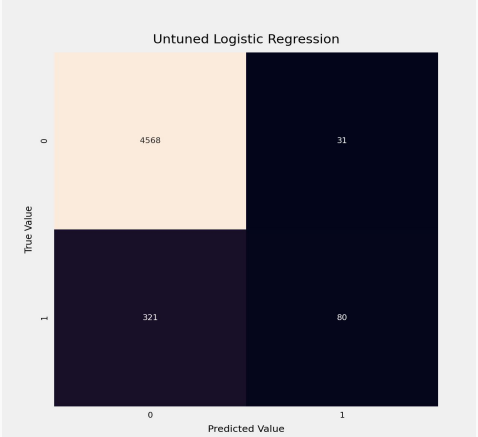
Baseline Model
Tuning Decision Boundary
Evaluation & Analysis

BASELINE MODEL - LOG REG

PERFORMANCE STATISTICS

		precision	recall	f1-score	support
	0.0	0.93	0.99	0.96	4599
	1.0	0.72	0.20	0.31	401
accuracy				0.93	5000
macro avg		0.83	0.60	0.64	5000
weighted avg		0.92	0.93	0.91	5000

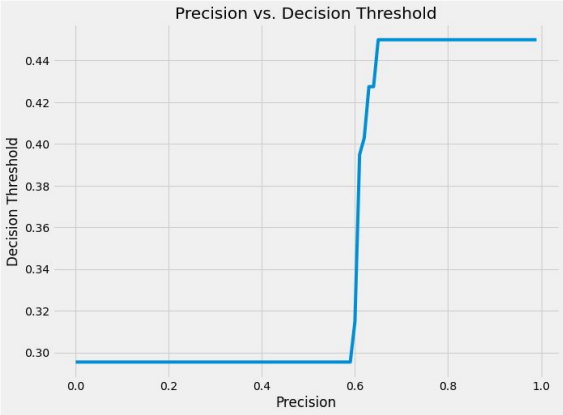
CONFUSION MATRIX



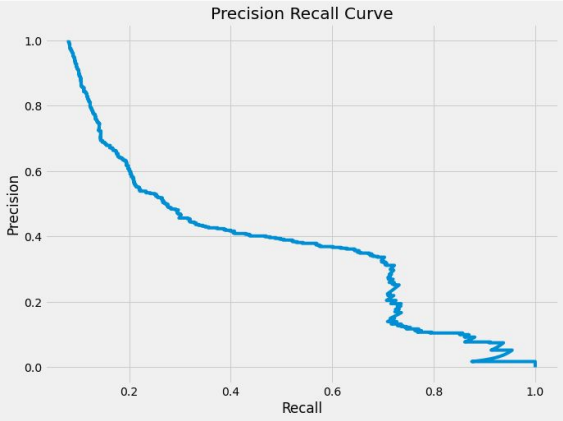
- Satisfactory accuracy (93%) and precision (72%) scores
- Inadequate, low recall score (20%) - implying that the model fails to classify nearly 4/5 of all defaulting clients
- Overall, model is extremely good at identifying that a client will NOT default, but in the process misses a lot of clients who will default
- Attention to optimizing the decision boundary is needed to reduce the amount defaulting clients missed by the model

TUNING DECISION BOUNDARY - LOG REG

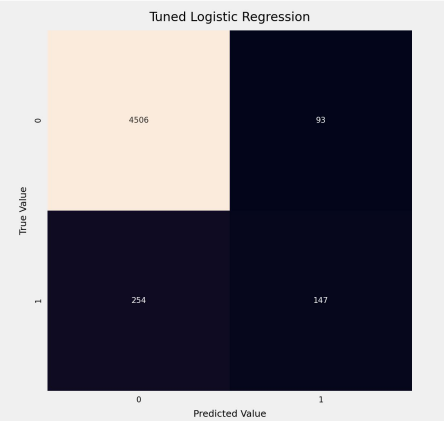
PRECISION VS DECISION



PRECISION VS RECALL



CONFUSION MATRIX



- Tuning Algorithm: *decrease threshold as long as precision does not fall below a certain value*
- Decreasing the decision threshold is equivalent to reducing the amount of risk the firm is willing to take on, we believe 0.15 is the optimal decision boundary
- After tuning, the model's recall was greatly improved while sacrificing a small amount of precision - allowing the firm to be extremely confident taking on less risk, in terms of identifying potential defaulters

PERFORMANCE STATISTICS

		precision	recall	f1-score	support
	0.0	0.95	0.98	0.96	4599
	1.0	0.61	0.37	0.46	401
accuracy				0.93	5000
macro avg		0.78	0.67	0.71	5000
weighted avg		0.92	0.93	0.92	5000

03.
FEED FORWARD NEURAL NETWORK

ML Algo Choice Explanation
Initial Model
Model Stability
Tuning Model Width
Tuning Decision Boundary

FEED FORWARD NEURAL NETWORK

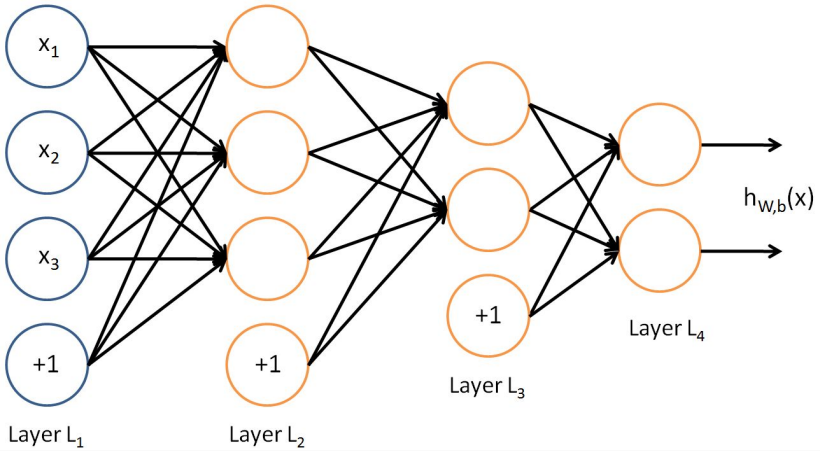


Fig. Example of a Feed-forward Neural Network with two hidden layers layers (L2,L3) and two output units in layer, L4
Source: deeplearning.stanford.edu

Decision to use Feed-forward Neural Network (FFNN) as our machine learning algorithm was based on the realization that the dataset contains a large number of continuous variables & possible interaction effects.

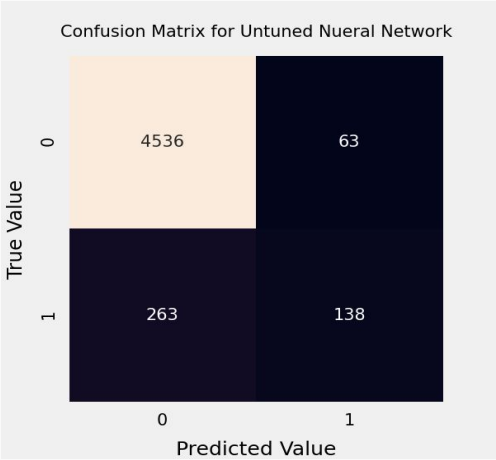
- FFNNs can be useful because they allow for a very large number of interaction effects without extensive testing
- FFNNs have the ability to capture nonlinear interaction effects that may exist between variables
- Cautious of FFNN's black-box nature, the cost of training, and the possibility of arriving at a subpar solution as a result of iterative parameter searching

INITIAL MODEL - FEEDFORWARD NN

PERFORMANCE STATISTICS

		precision	recall	f1-score	support
	0.0	0.93	0.99	0.96	4599
	1.0	0.76	0.19	0.31	401
accuracy				0.93	5000
macro avg		0.85	0.59	0.64	5000
weighted avg		0.92	0.93	0.91	5000

CONFUSION MATRIX



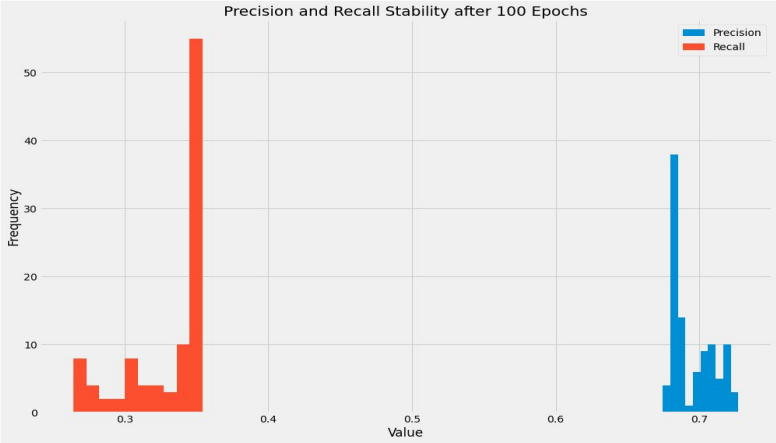
- Single hidden layer with a relu activation function & single output node of with a sigmoid activation function
- Started with the same number of hidden nodes as input features, and a loss based on binary cross entropy
- Model performance is similar to the original logistic regression

MODEL STABILITY - FEEDFORWARD NN

PERFORMANCE STATISTICS

		precision	recall	f1-score	support
	0.0	0.95	0.99	0.97	4599
	1.0	0.69	0.34	0.46	401
accuracy				0.93	5000
macro avg		0.82	0.67	0.71	5000
weighted avg		0.92	0.93	0.92	5000

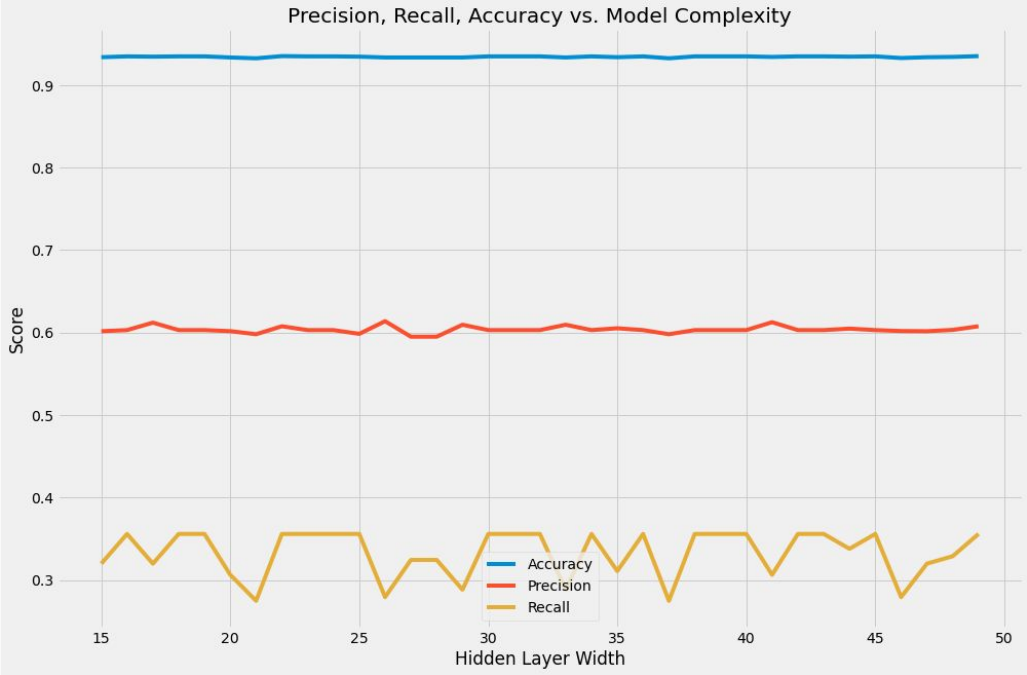
DISTRIBUTION OF PRECISION & RECALL SCORES



- Ensured model stability by training Feed-forward neural network 100 times over 100 epochs
- Recall scores are on average much higher than the recall scores from the logistic regression
- Extremely promising - almost outperforming our tuned logistic regression model

TUNING MODEL WIDTH - FEEDFORWARD NN

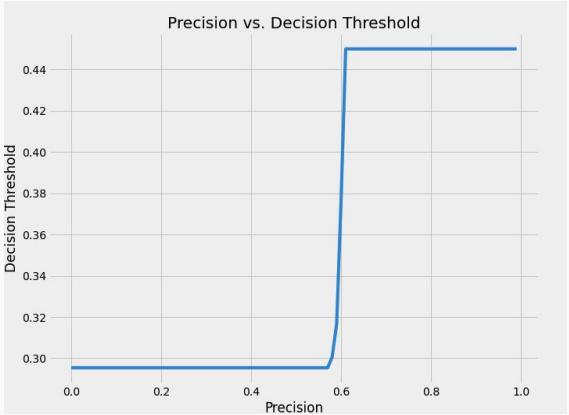
PRECISION, RECALL, ACCURACY VS MODEL COMPLEXITY



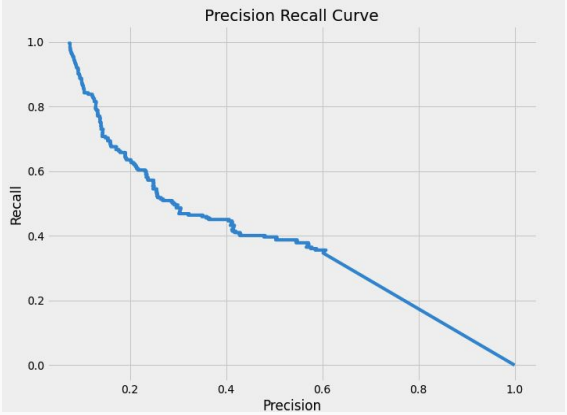
- This graph provides the change of precision, recall, and accuracy, as the number of nodes in the hidden layer
- Inferred: No tangible benefit to increasing the model complexity past our current point

TUNING DECISION BOUNDARY - FEEDFORWARD NN

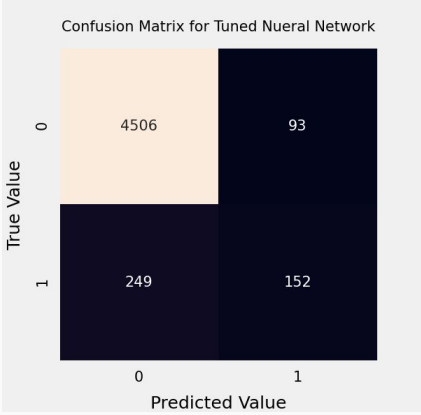
PRECISION VS DECISION



PRECISION VS RECALL



CONFUSION MATRIX



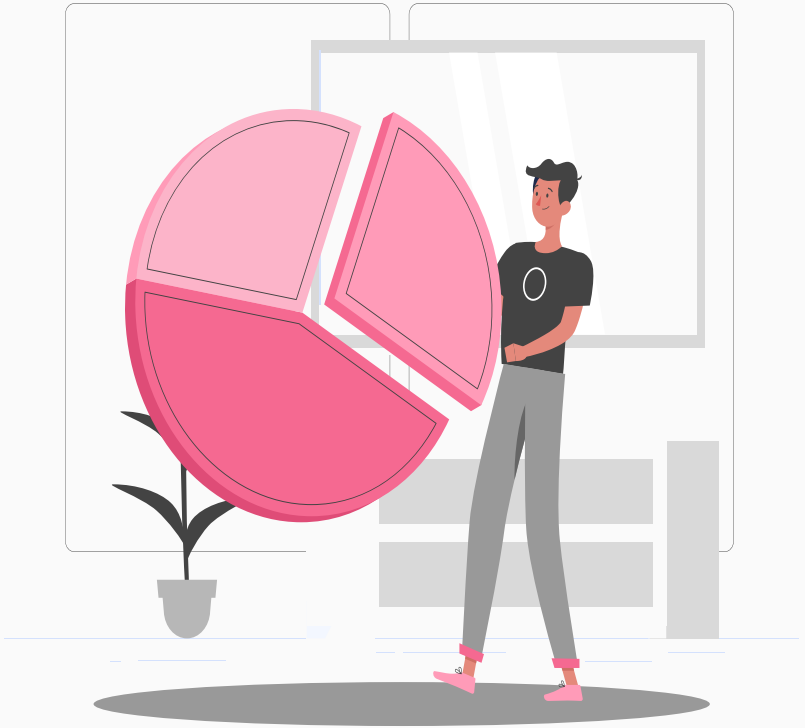
- Precision recall curve is more preferable than the logistic regression model
- The tuning precision of 0.55 seems to be a good sacrifice as the precision recall curve is quite steep
- Beyond this point, improving recall requires a much greater sacrifice in precision and it will not be in the firm's interest to go beyond this decision boundary

PERFORMANCE STATISTICS

		precision	recall	f1-score	support
	0.0	0.95	0.98	0.96	4599
	1.0	0.61	0.37	0.46	401
accuracy				0.93	5000
macro avg		0.78	0.67	0.71	5000
weighted avg		0.92	0.93	0.92	5000

04.
CONCLUSIONS

Evaluation & Analysis
Future Approval Decisions
Bias Towards Customers



MAKING AND EXPLAINING FUTURE DECISIONS

- A very simple approach can be used to explain our decisions: compare a given customer's risk to the mean risk a random customer presents (0.213).
    1. Hold all values at their mean
    2. Change a single factor to the customers true value
    3. Compare the change in risk
- This approach provides us a solid idea to why we are rejecting or accepting a credit application.

We recommend using our model as an initial screening tool for the firm.

- Our model displays risk-taking behavior as it systematically underestimates when a customer will default.
- **If a customer can't pass this initial screening there is a high possibility that they will default in the future**
    - We arrive at this conclusion based on our precision values
- We **do not recommend our model as the final screening** as it misses many people who will default

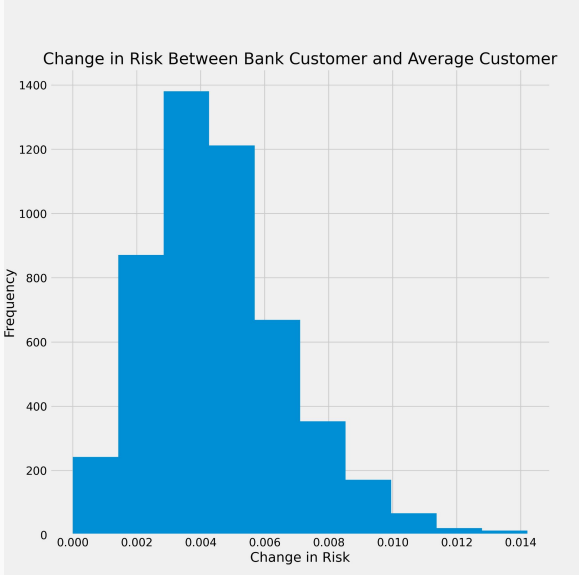
	Person45	Person1
tot_credit_debt	1.000000e+00	0.514523
avg_card_debt	4.437685e-03	0.137980
credit_age	3.770228e-01	0.372324
credit_good_age	2.526338e-01	0.257978
card_age	3.722641e-01	0.383290
mortgages_past_due_6_months_num	2.138428e-01	0.213843
credit_past_due_amount	3.349119e-02	0.033491
inq_12_month_num	1.915515e-01	0.200272
card_inq_24_month_num	1.846877e-01	0.196563
uti_card	2.094540e-01	0.210985
uti_50plus_pct	2.102203e-01	0.211625
uti_max_credit_line	2.107638e-01	0.210582
uti_card_50plus_pct	2.099449e-01	0.211877
ind_acc_XYZ	2.188845e-01	0.209320
rep_income	5.187532e-25	0.000000
Default_ind	2.143294e-01	0.213067
FL	2.130757e-01	0.213076
SC	2.138326e-01	0.213833
LA	2.152982e-01	0.215298
GA	2.134444e-01	0.213444
MS	2.144244e-01	0.214424
NC	2.138411e-01	0.213841
non_mtg_acc_past_due_12_months_num==2.0	2.129159e-01	0.212916
non_mtg_acc_past_due_12_months_num==1.0	2.139596e-01	0.213960
non_mtg_acc_past_due_12_months_num==3.0	2.139562e-01	0.213956
non_mtg_acc_past_due_12_months_num==4.0	2.133930e-01	0.213393
non_mtg_acc_past_due_6_months_num==1.0	2.139547e-01	0.213955
non_mtg_acc_past_due_6_months_num==2.0	2.130403e-01	0.213040
card_open_36_month_num==1.0	2.138993e-01	0.213899
card_open_36_month_num==2.0	2.130323e-01	0.213032
auto_open_ 36_month_num==1.0	2.139616e-01	0.213962

DO WE BIAS TOWARDS PERVIOUS BANK CUSTOMERS?

A similar approach is used to determine if being a previous customer has any benefit in our model. We used the mean values for all other variables and changed only the indicator for if someone owns a bank account.

Average Customer Risk: **0.2139**
Average Risk for Previous Customer: **0.2069**

- **Previous bank customer enjoy a marginal benefit**, but we can take this one step further and generate a whole distribution to measure this benefit
- This distribution is generated by manipulating the previous customer indicator for 5,000 customers in the test dataset
- **Bank customers are seen as 0% to 1.4% less risky than non-bank customers when it comes defaulting**, center is around 0.4% and skewed to the right
- **The bias towards bank customers is unlikely to cause a change in approval decision**, and may reflect patterns found in the real world



THANKS

A special thank you to the **UGA Department of Statistics, Wells Fargo, & Dr. Mohamad Al Lawati** for organizing this challenging, learning opportunity.

Does anyone have any questions?