Analysis of Non-Payment Among Canadian Households

Oscar Cueva

Business Analytics Capstone Project

Centennial College

High-Level Project Details

Description – *The What*

Business Case – *The Why*

Business Objective – *The Goal*

Overview of Findings – *The Result*

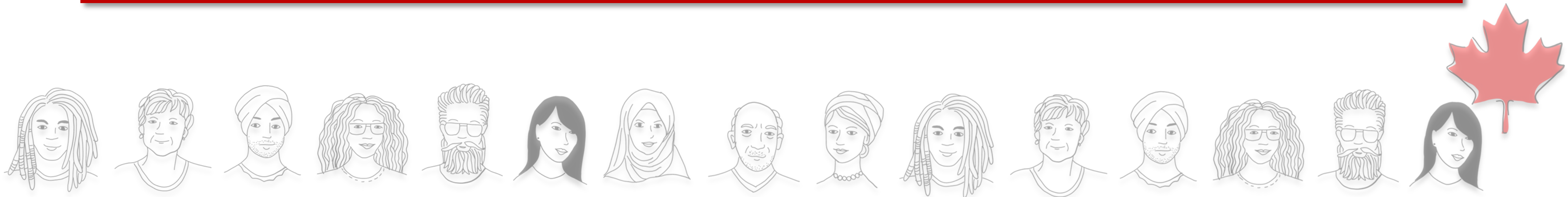
Canadians Skipping or Delaying Payments

- 42 Machine Learning models were built to answer:

Will a Canadian Household skip (or delay) a non-mortgage payment?

- Businesses face a common but complicated task:

Assessing clients' probability of non-payment and associated costs

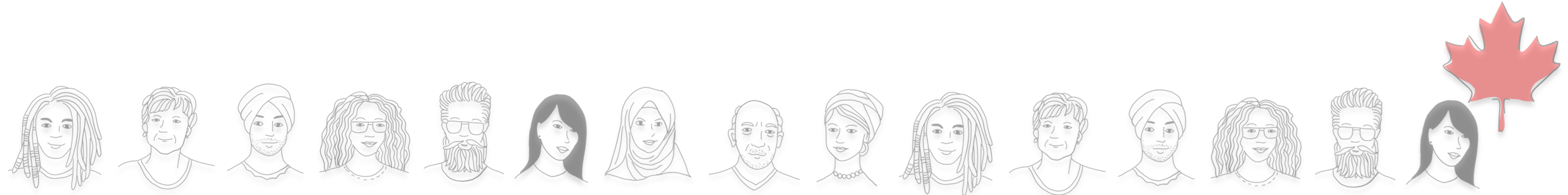


Efficient Predictor & Actionable Insights

- This project has two short-term objectives:

Identify the households with the highest and lowest probabilities of skipping (or delaying) a payment

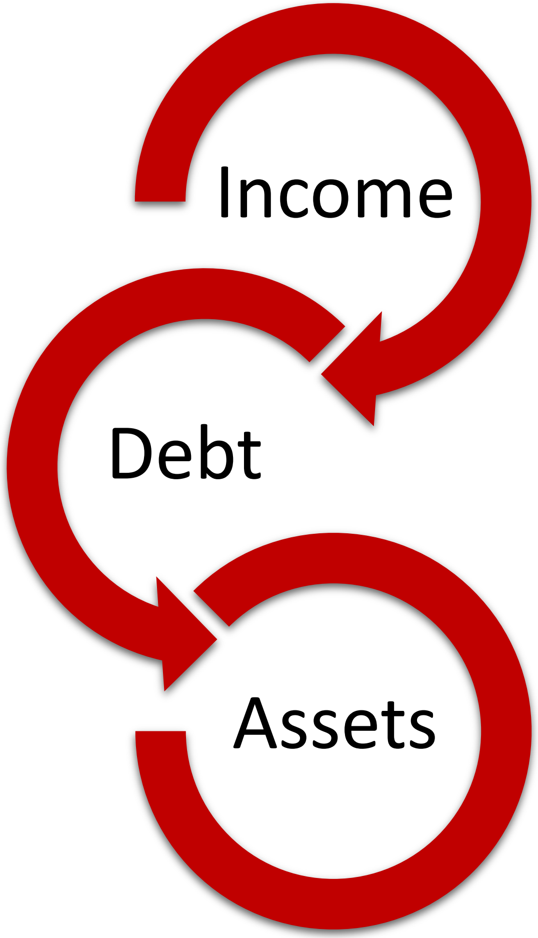
Identify the main driving factors of such probability to generate actionable insights



Drivers: Interaction of Income, Debt & Assets

- Among various other factors, the probability of non payment is driven by:

<u>Interaction</u> of multiple components of a household's net worth*.



*This finding has to be researched further

Analytics Process

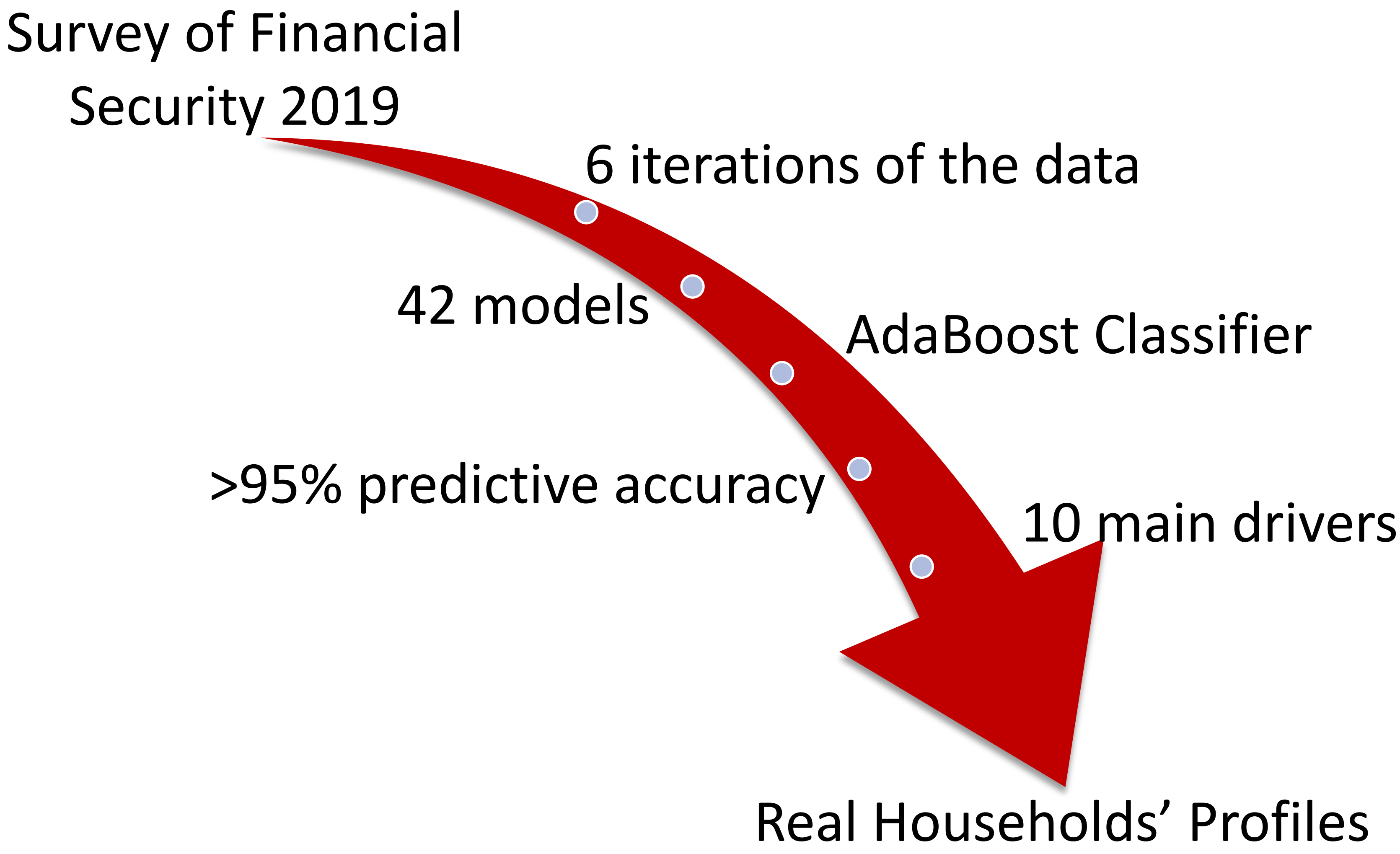
Data

Models

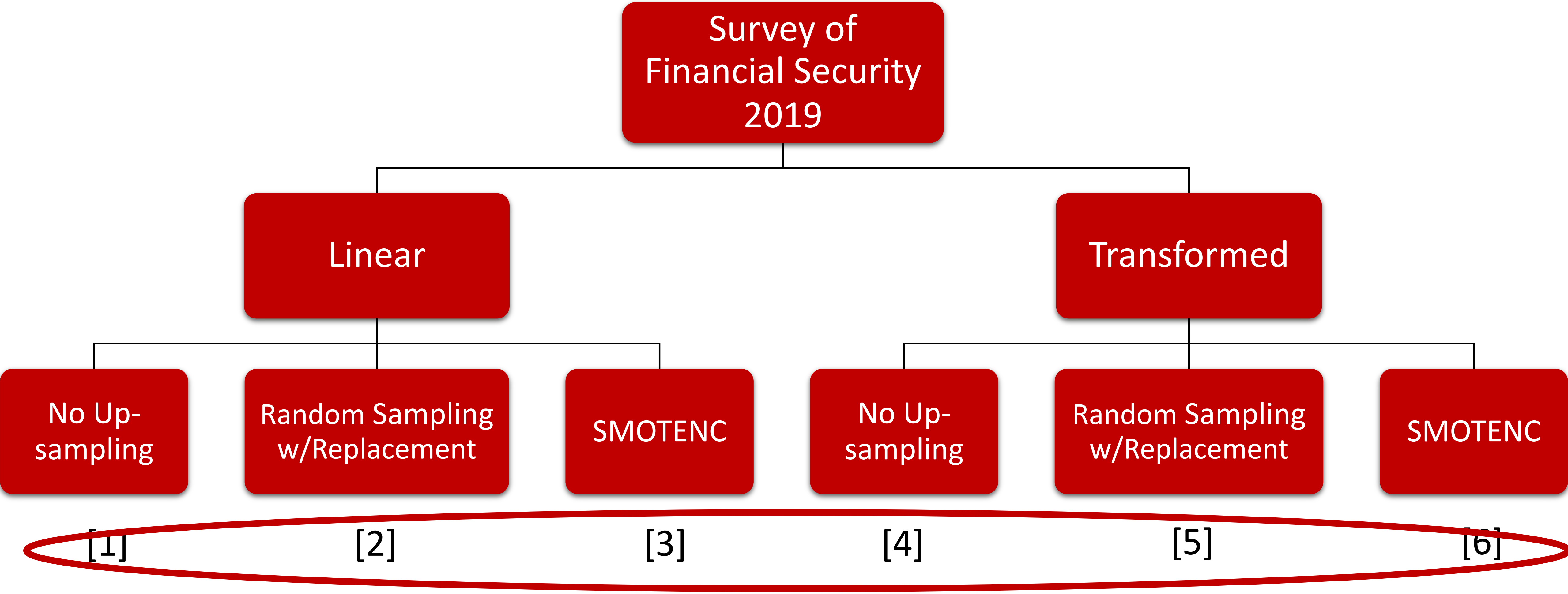
Accuracy & Measures of Success

Households' Profiles

The Process of Turning Data into Real People



Real Data – Real People – Real Challenges



Models: Classifiers(Tree/Regression/Neural Network)

- Full Classification Trees
- Random Forest Classifiers
- Logistic Regression
- AdaBoost Classifier
- Feature Selection: Forward / Backward / Stepwise
- Neural Network

AUCROC: 96.15%

		Base			Trans		
		AUCROC	Recall	Accuracy	AUCROC	Recall	Accuracy
NO-SAMP	Full Tree	55.72%	18.55%	87.08%	55.79%	18.55%	87.72%
	RFC	79.60%	0.00%	92.19%	79.63%	0.00%	92.19%
	AdaBoost	68.53%	8.69%	91.52%	70.01%	7.25%	91.36%
	Logistic	76.12%	70.72%	67.17%	80.66%	72.67%	73.39%
	NN	80.22%	64.26%	77.65%	80.62%	65.05%	78.73%
UP-SAMP	Full Tree	54.63%	15.58%	54.63%	54.38%	14.80%	54.38%
	RFC	78.80%	75.80%	72.09%	78.71%	75.88%	72.11%
	AdaBoost	70.85%	9.99%	70.85%	68.09%	8.27%	53.03%
	Logistic	75.05%	68.25%	68.92%	79.61%	73.84%	72.86%
	SW	78.79%	72.49%	72.02%	80.89%	74.09%	73.75%
	BWD	--	--	--	--	--	--
	FW	78.53%	72.61%	72.04%	80.89%	74.09%	75.76%
	NN	79.74%	67.39%	71.99%	79.33%	72.75%	71.23%
SYN	Full Tree	78.99%	68.49%	78.99%	83.36%	75.95%	83.36%
	RFC	89.30%	85.42%	80.61%	91.85%	91.60%	83.45%
	AdaBoost	95.50%	78.94%	88.34%	96.15%	84.49%	91.28%
	Logistic	75.80%	68.79%	69.01%	89.44%	79.31%	80.83%
	SW	84.35%	78.50%	77.15%	89.19%	80.39%	83.89%
	BWD	--	--	--	--	--	--
	FW	84.04%	80.66%	77.80%	89.19%	80.39%	83.89%
	NN	89.61%	72.39%	80.07%	95.54%	84.27%	88.69%

AUCROC is Best Fit Project Business Case

- What is AUCROC?

Area Under the ROC curve

ROC plots the relationship between True Positive to False Positive predictions at different classification thresholds.*

- Why AUCROC?

The AUCROC measure how efficient a model is at predicting true positives

*Developers Google, 2022

Customer Profiling Found Diversity

	HIGH AND LOW PROBABILITY OF MISSING OR SKIPPING A PAYMENT						
Concept	Hi-Inc Hi-Proba	Lo-Inc Hi-Proba	Hi-Inc Lo-Proba	Lo-Inc Lo-Proba	Max	Min	Mean
Government Transfer	1435.0	18957.0	500.0	325.0	49326.0	0.0	10843.0
After-Tax Income	152632.0	18957.0	176975.0	46075.0	232120.0	-26000.0	71199.0
Other Non-Financial Assets	24389.0	365.0	9500.0	5000.0	184793.0	175.0	17702.0
Total Money in Banks	1666.0	55.0	4750.0	20000.0	221297.0	-3200.0	12520.0
Total Credit Limit (CC)	17806.0	0.0	15000.0	0.0	88534.0	0.0	15228.0
Total Debt on Vehicles	39639.0	464.0	82500.0	15500.0	96432.0	0.0	18566.0
Net Worth of Family	1031923.0	1812.0	774400.0	500250.0	2831148.0	-910500.0	485644.0
Total Debts of Family	270398.0	0.0	389100.0	5250.0	636230.0	0.0	107091.0
Age of Major Income Earner	49.0	62.0	39.0	69.0	84.0	19.0	50.0
Total CC and Installment Debt	5744.0	0.0	4100.0	0.0	24162.0	0.0	2919.0

Business Insights

Insights

Recommendations

Limitations

Insights – The Importance of Financial Ratios

The relationship *or ratio* of assets to liabilities, and income to debt seems to be important.

- High and low income households were found on both sides of the probability spectrum
- The relationship of assets to liabilities seems to remain relevant
- This finding is specially important to prevent biased assessments

Insights – The Importance of Payment History

- Complementary findings suggest that:

Payment History plays an important role to predict future behavior

Collateral reduces the risk of non-payment – real estate, vehicles, cash

Recommendations

- If looking to mitigate the probability of skipped payments:

Concentrate on households with healthy financial ratios

Avoid pritoritizing income or wealth alone

Look at past payment history

Look at customers' potential collateral

Limitations

- The results of this project are based upon household data
- The predictions from this project should be applied solely on households
- There is an exclusion of the 15% wealthiest Canadian households as measured across multiple dimensions of wealth
- This model should not be used to households with a net worth that exceeds $2.8 million CAD, or with an annual after-tax income greater than $232,000 CAD

Next Steps - Clustering Analysis

- Investigate the potential interaction among variables
- Pre and post-pandemic clustering analysis to compare how groups of interest behaved prior and after the economic shock caused by the COVID-19 pandemic

Questions

Statistics Canada Disclaimer

“This analysis is based on Statistics Canada’s Survey of Financial Security Public Use Microdata, 2019, which contains anonymous data collected in the Survey of Financial Security. All computations on these microdata were prepared by Oscar Cueva. The responsibility for the use and interpretation of these data is entirely that of the author.”

Thank You!

References

- *Business cycle dating committee announcement July 19, 2021*. NBER. (2021). Retrieved August 12, 2022, from https://www.nber.org/news/business-cycle-dating-committee-announcement-july-19-2021
- *How global labor markets are recovering from covid-19*. Deloitte Insights. (2022). Retrieved August 12, 2022, from https://www2.deloitte.com/us/en/insights/economy/issues-by-the-numbers/impact-covid-19-labor-market-globally.html
- *News release*. Gross Domestic Product, Third Quarter 2020 (Advance Estimate) | U.S. Bureau of Economic Analysis (BEA). (2020). Retrieved August 12, 2022, from https://www.bea.gov/news/2020/gross-domestic-product-third-quarter-2020-advance-estimate
- Remes, J., Manyika, J., Smit, S., Kohli, S., Fabius, V., Dixon-Fyle, S., & Nakaliuzhnyi, A. (2021, December 15). *The consumer demand recovery and lasting effects of COVID-19*. McKinsey & Company. Retrieved August 12, 2022, from https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/the-consumer-demand-recovery-and-lasting-effects-of-covid-19
- Rockefeller, B. (2022, May 20). *Global Supply Chain Pressure index: May 2022 update*. Liberty Street Economics. Retrieved August 12, 2022, from https://libertystreeteconomics.newyorkfed.org/2022/05/global-supply-chain-pressure-index-may-2022-update/
- Statistics Canada. (2020, December 22). *Survey of Financial Security*. Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1252634

Appendix I. Government Transfers: Main Driver

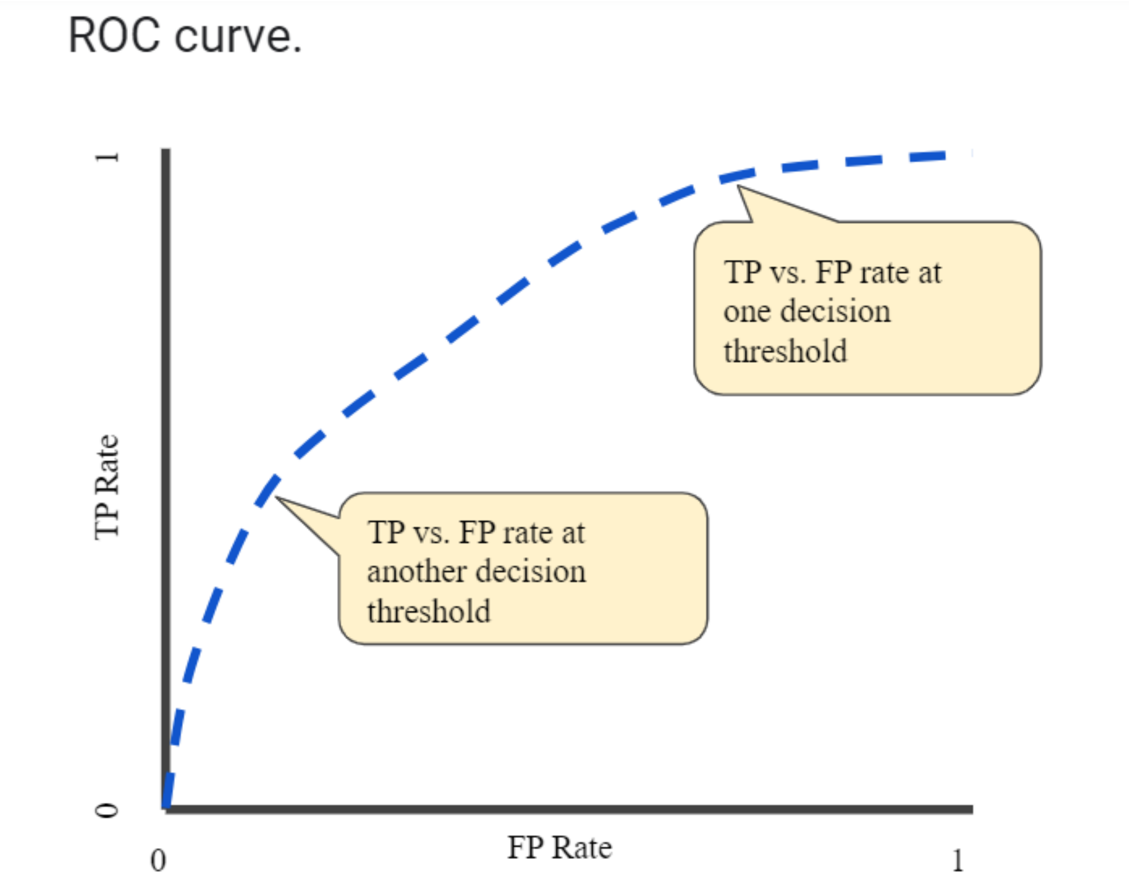
```
Confusion Matrix (Accuracy 0.9128)

       Prediction
Actual     0    1
     0 3996   79
     1  632 3443

AdaBoost Decision Tree Classifier (Trans-Syn) AUCROC: 0.9615
```

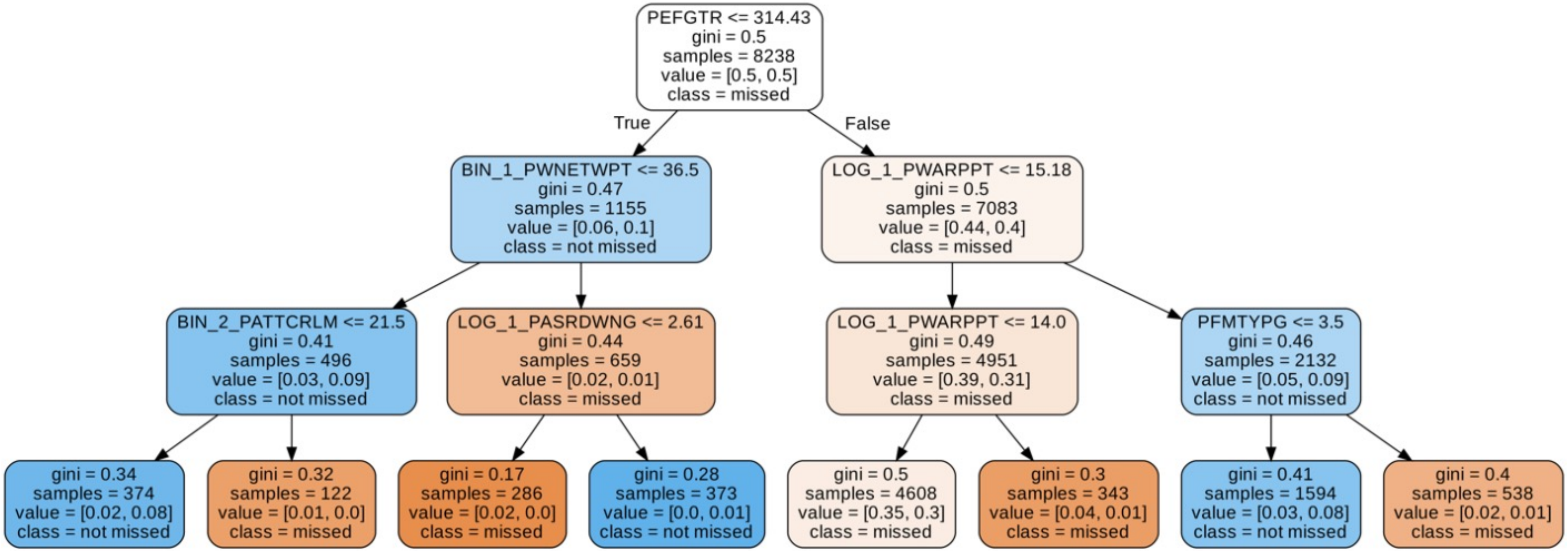
Government Transfers → After-Tax Income → Total Money in Banks → Total Credit Card and Debt Installment → Total Credit Card Limit → Total Debt on Vehicles → Age of Major Income Earner

Appendix II. AUCROC



Developers Google, 2022.
https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Appendix III. Model: Tree Visualization

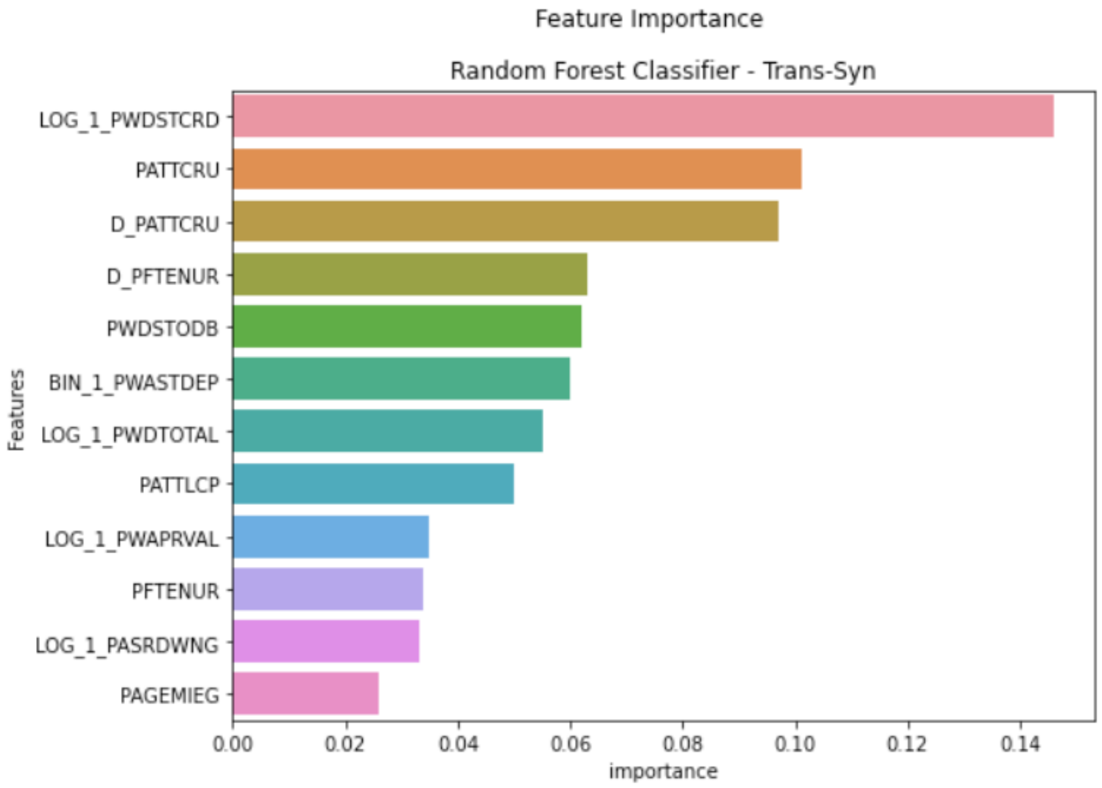


Appendix IV. Secondary Model Metrics

```
Confusion Matrix (Accuracy 0.8345)

       Prediction
Actual    0    1
     0 3068 1007
     1  342 3733

Random Forest Classifier (Trans-Syn) AUCROC: 0.9185
Training Accuracy: 0.8491
```



Appendix V. Secondary Model Profiling

HIGH AND LOW PROBABILITY OF MISSING OR SKIPPING A PAYMENT								
	Concept	Hi-Inc Hi-Proba-	Hi-Inc Hi-Proba	Hi-Inc Lo-Proba	Lo-Inc Lo-Proba	Max	Min	Mean
PEFATINC	After Tax Income	109299.0	65413.0	140375.0	33125.0	232120.0	-26000.0	71199.0
PWDSTCRD	Total CC and Installment Debt	1982.0	1641.0	0.0	0.0	24162.0	0.0	2919.0
PATTCRU	CC Balance Paid on Time?	3.0	3.0	4.0	4.0	NaN	NaN	NaN
D_PATTCRU	Paid the Minimum or Less?	1.0	1.0	0.0	0.0	NaN	NaN	NaN
D_PFTENUR	Owns a Home with No Mortgage?	0.0	0.0	1.0	1.0	NaN	NaN	NaN
PWASTDEP	Total Money in Banks	1511.0	-289.0	45000.0	10000.0	221297.0	-3200.0	12520.0
PWDSTODB	Total of other Debt	0.0	782.0	0.0	0.0	8050.0	0.0	674.0
PWDTOTAL	Total Debts of Family	209731.0	125979.0	0.0	9500.0	636230.0	0.0	107091.0
PATTLCP	Line of Credit Balance Paid on Time?	2.0	0.0	4.0	0.0	NaN	NaN	NaN
PWAPRVAL	Value of Principal Residence	551464.0	107816.0	950000.0	95000.0	1257136.0	0.0	237959.0
PAGEMIEG	Age of Major Income Earner	54.0	37.0	69.0	59.0	84.0	19.0	50.0

Appendix VI. Best Neural Network Metrics

```
Initial bias: 0.0
Weight for class 0: 1.00
Weight for class 1: 1.00

Restoring model weights from the end of the best epoch: 403.
Epoch 418: early stopping

Results:
    TN - FN: 3794 / 641
    FP - TP: 281 / 3434

    Validation Accuracy: 0.8869
    Validation AUC_ROC: 0.9554
    Validation Recall: 0.8427
    Precision: 0.9244
```