Empowering Financial Inclusion:

Credit default prediction through alternative data integration

Marian Ilagan

EXECUTIVE SUMMARY

Leveraging alternative data sources can offer a more comprehensive view of borrowers' creditworthiness, especially for the credit-disadvantaged segment. In this project we explored to incorporate the use of alternative data in predictive models for credit default.

TOPICS OF DISCUSSION

Business Opportunity

Project Proposal

Conclusion

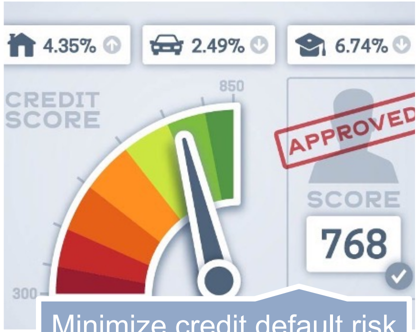
1

BUSINESS OPPORTUNITY

Credit score is one of the main considerations in loan approvals

People apply for loans for different reasons

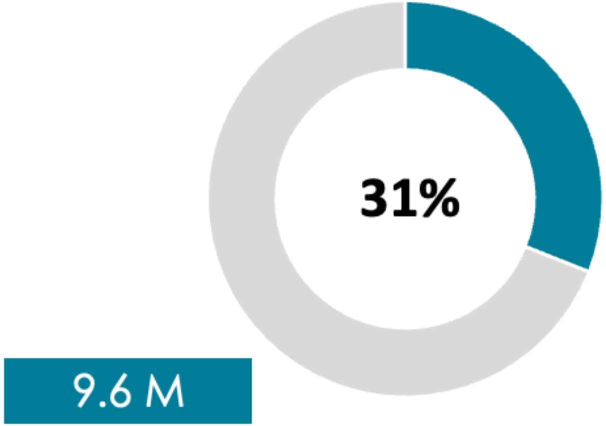
Lenders get profit from loans, only if they are paid back



Minimize credit default risk by assessing client's financial capability

Dependency on credit score creates a challenge for credit-disadvantaged customers

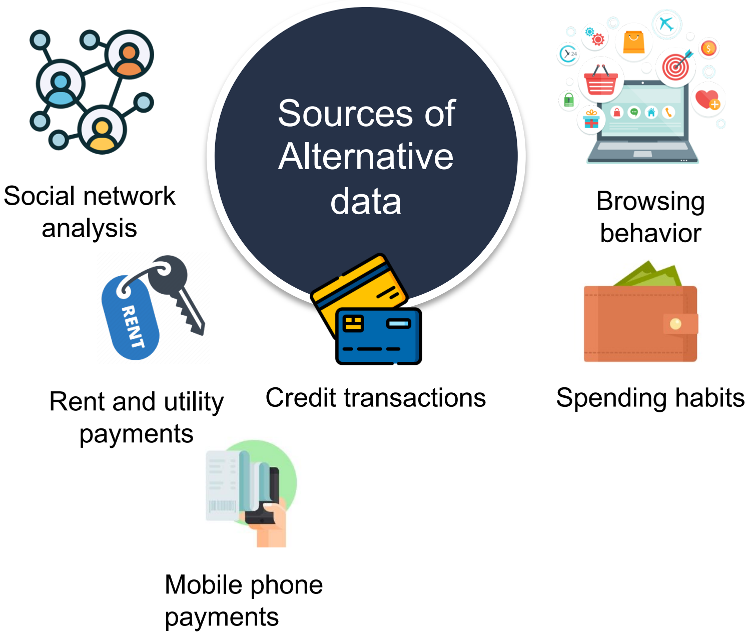
Percentage of Canada's adult population that is underserved or unserved[1].



- Unable to access financial services because of little to no credit history
- Low credit score often leads to higher interest rates or rejection
- 90% of underserved migrated to credit served within two years
- Higher level of inquiry implies demand is not being met

1 TransUnion: More than 9 Million Canadians are Either Credit Unserved or Underserved

Alternative data can provide fuller picture of financial capacity



Data that is not usually part of credit reports

↓

Enables lenders to make credit more available

↓

Increase company's market share by tapping on the credit underserved

2

PROJECT PROPOSAL

OBJECTIVES

Business Goal

- Understand the overall financial capacity of underserved consumers
- Generate new credit applications and increase customer base

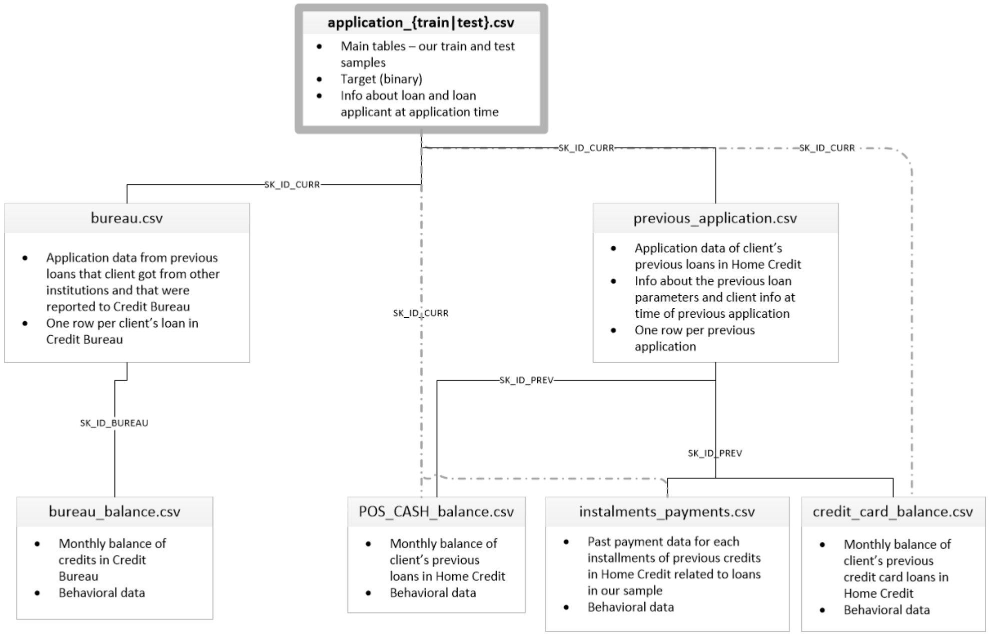
Analytics Goal

- Develop a default risk prediction model that will ensure underserved customers who are capable of repayment are not rejected.

Research Questions

- What features from alternative data are important for default prediction?
- Which model can provide the most optimal solution?

DATA SOURCE



Kaggle: Home Credit Default Risk

- ❖International consumer finance provider
- ❖Provides positive and safe borrowing experience to people with little or no credit history
- ❖Makes use of alternative data to predict client's repayment abilities
- ❖Cash loans, point-of-sale loans, credit cards

DATA DESCRIPTION

	application_df	previous_df	bureau_df	bureaubal_df	POS_df	prev_cc_df	installments_df
rows	307511	1670214	1716428	27299925	10001358	3840312	13605401
columns	122	37	17	3	8	23	8

Credit related

- Annuity, credit score from external sources, number of inquiries, type of loans, purpose of loan

Demographic

- age, gender, highest level of education, occupation, count of children, count of family members

Transactions

- credit card drawings, point-of-sale, cash loans

Social network

- Number of defaults within the client's social circle

Living situation

- housing situation, rating of the region, building information of where the client lives, owns a car

Payment history

- Number of days of late payment, number of remaining installments

PROJECT AND DATA SCOPE

Assumptions	• All variables are free and acceptable to use • Variables are legally available in Canada
Exclusions	• First time applicants • Clients with no data in credit bureau
Risk and Dependencies	• Skewed data • Imbalanced dataset • Large dataset

PROCESS FLOW

Initial Data Cleaning and Preparation

- Removed data for exclusion
- Analyzed supporting tables – Feature engineering
- Prepared final table

Data Exploration, Preprocessing

- Missing values
- Multicollinearity
- Identify Outliers
- Skewed data
- Irrelevant values

Modeling

- Feature Selection
- Oversampling and under sampling
- Models
- Optimization

Best Model

- Choose model based on ROC, F1 score, Accuracy
- Feature Importance

MODEL PERFORMANCES

	ROC-AUC	F1 Score	Accuracy
Random Forest	0.69	0.26	0.77
Stochastic Gradient Descent	0.69	0.26	0.68
Logistic Regression	0.69	0.26	0.64
Xgboost	0.69	0.20	0.86
AdaBoost	0.67	0.20	0.88
Decision Tree	0.55	0.19	0.75

BEST MODEL RESULTS

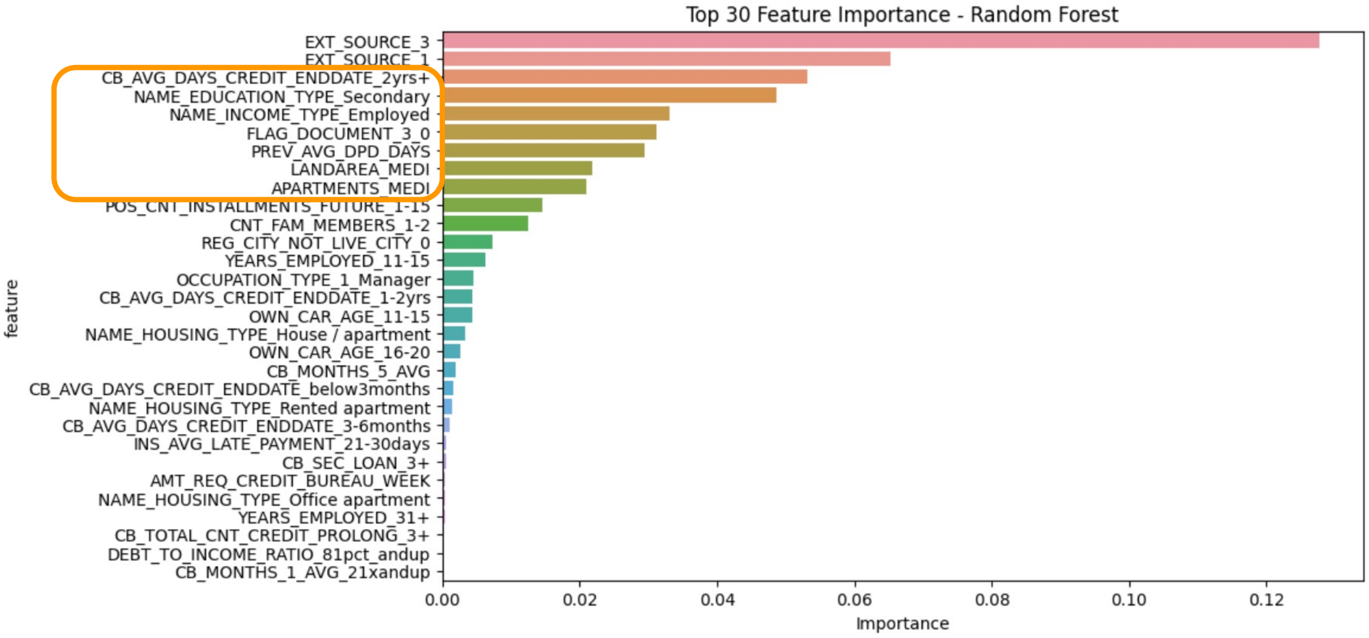
	Prediction	
Actual	Not defaulted	Defaulted
Not defaulted	**12151**	**2783**
Defaulted	**941**	**641**

Recall: Captures **41%** of the defaulted clients

Precision: **19%** of the time the model is right in identifying default.

Accuracy: The model is correct **77%** of the time.

FEATURE IMPORTANCE



3

CONCLUSION & RECOMMENDATION

RECOMMENDATIONS

Analytics Team

- Tailor, refit and optimize the model to Canadian market
- Improve model scores:
  - Create high-quality features that can better predict default
  - Modify thresholds
- Create models for clients without credit history

Management

- Review regulatory laws about using alternative data
- Incorporate alternative data when evaluating the 5Cs of the applicant:
  - NAME_INCOME_TYPE (Employed or Unemployed), OCCUPATION TYPE – Capacity to pay
  - Previous average DPD (days past due) days – Character or repayment behavior
- Explore the use of explainable AI (e.g. SHAP)

CONCLUSION

Analytics Goal

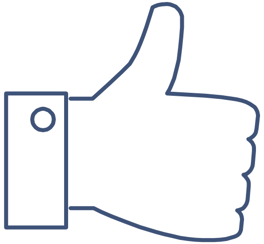
- Incorporated alternative data from Home Credit
- Predicted underserved customers who are capable of repayment

Research Questions

- Best Model: **Random Forest** 70% ROC-AUC
- Top 3 features: CB_AVG_CREDIT_ENDDATE 2yrs+, NAME_EDUCATION_TYPE (Secondary), NAME_INCOME_TYPE (Employed)

Business Goal

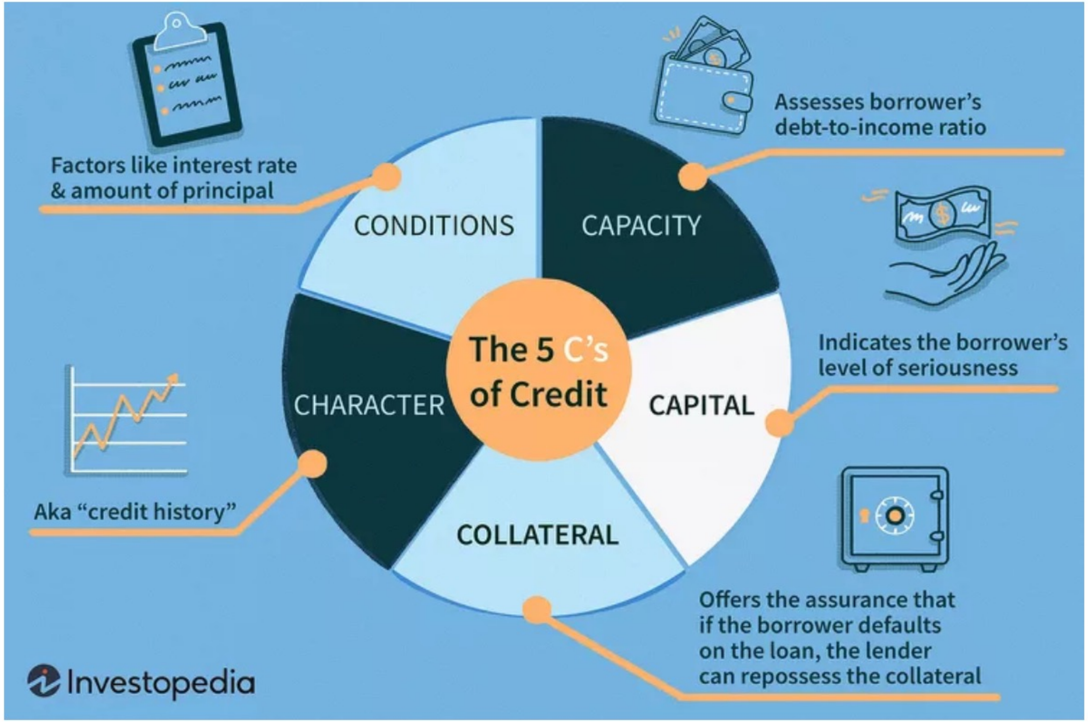
- Understood client's financial capabilities
- Make more informed lending decisions
- Empower Financial Inclusion



THANK YOU!

Any questions?

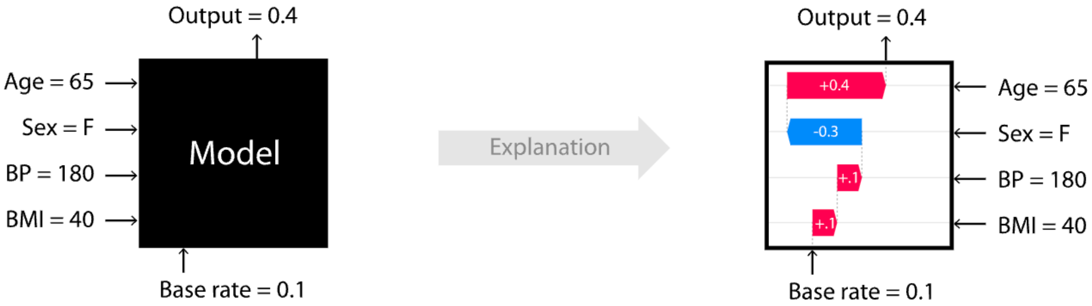
APPENDIX


Factors like interest rate
& amount of principal
Assesses borrower's
debt-to-income ratio
CONDITIONS
CAPACITY
The 5 C's
of Credit
CHARACTER
CAPITAL
Indicates the borrower's
level of seriousness
COLLATERAL
Aka "credit history"
Offers the assurance that
if the borrower defaults
on the loan, the lender
can repossess the collateral
Investopedia

SGD / Regression – Feature Importance

	coef	odds_ratio
CB_AMT_CREDIT_SUM_OVERDUE_76-100k	1.590064	4.904062
AMT_REQ_CREDIT_BUREAU_WEEK	0.598804	4.701518
NAME_INCOME_TYPE_Employed	1.109093	3.031609
CB_SEC_LOAN_0	0.864272	2.373278
DEF_30_CNT_SOCIAL_CIRCLE	0.303935	2.025294
OCCUPATION_TYPE_8_Operators_Assemblers	0.546727	1.727589
NAME_EDUCATION_TYPE_Secondary	0.541917	1.719299
OWN_CAR_AGE_21-25	0.501692	1.651513
FLAG_OWN_CAR_N	0.437463	1.548773
CB_AVG_DAYS_CREDIT_ENDDATE_2yrs+	0.297943	1.347085

SHAP



https://github.com/shap/shap

PCA

Columns after PCA	Description
AMT_REQ_CREDIT_BUREAU_HOUR	Number of inquiries 1 hr before the application
AMT_REQ_CREDIT_BUREAU_WEEK	Number of inquiries 1 wk before the application
APARTMENTS_MEDI	Normalized information about the building (median apartment size)
CB_AMT_CREDIT_SUM_OVERDUE_100+	Amount overdue in Credit bureau
CB_AMT_CREDIT_SUM_OVERDUE_26-50k	Amount overdue in Credit bureau
CB_AMT_CREDIT_SUM_OVERDUE_51-75k	Amount overdue in Credit bureau
CB_AMT_CREDIT_SUM_OVERDUE_76-100k	Amount overdue in Credit bureau
CB_AMT_CREDIT_SUM_OVERDUE_no_overdue	Amount overdue in Credit bureau
CB_AVG_DAYS_CREDIT_ENDDATE_1-2yrs	AVG remaining duration of CB credit (in days) at the time of application in Home Credit
CB_AVG_DAYS_CREDIT_ENDDATE_2yrs+	AVG remaining duration of CB credit (in days) at the time of application in Home Credit
CB_AVG_DAYS_CREDIT_ENDDATE_3-6months	AVG remaining duration of CB credit (in days) at the time of application in Home Credit
CB_AVG_DAYS_CREDIT_ENDDATE_7-12months	AVG remaining duration of CB credit (in days) at the time of application in Home Credit
CB_AVG_DAYS_CREDIT_ENDDATE_below3months	AVG remaining duration of CB credit (in days) at the time of application in Home Credit
CB_MONTHS_1_AVG_21xandup	AVG times the client has a DPD within 30days (credit bureau)
CB_MONTHS_1_AVG_none	AVG times the client has a DPD within 30days (credit bureau)
CB_MONTHS_5_AVG	AVG times the client has a DPD more than 120+ days or written off
CB_SEC_LOAN_0	Number of secured loans
CB_SEC_LOAN_3+	Number of secured loans
CB_TOTAL_CNT_CREDIT_PROLONG_0	count of times the Credit Bureau credit prolonged
CB_TOTAL_CNT_CREDIT_PROLONG_3+	count of times the Credit Bureau credit prolonged
CC_AVG_AMT_DRAWINGS	Average amount of credit card drawings
CC_PCT_CASH_DRAWINGS	Percentage of credit card cash drawings
CF_CB_MONTHS_0_AVG	Average number of times the client paid on time (credit bureau)
CF_OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) - payment behavior
CNT_ACTIVE_LOANS	Number of active loans
CNT_FAM_MEMBERS_1-2	Count of family members
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
DEBT_TO_INCOME_RATIO_21-40pct	Debt to income ratio
DEBT_TO_INCOME_RATIO_41-60pct	Debt to income ratio
DEBT_TO_INCOME_RATIO_61-80pct	Debt to income ratio
DEBT_TO_INCOME_RATIO_81pct_andup	Debt to income ratio
DEBT_TO_INCOME_RATIO_none	Debt to income ratio
DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)

PCA

Columns after PCA	Description
EXT_SOURCE_1	Normalized credit score from external source
EXT_SOURCE_2	Normalized credit score from external source
EXT_SOURCE_3	Normalized credit score from external source
FLAG_DOCUMENT_3_0	client did not submit document 3
FLAG_OWN_CAR_N	Client has no car
FLAG_WORK_PHONE_0	client has no work phone
INS_AVG_LATE_PAYMENT_21-30days	average number of days client has late payment (installments)
INS_AVG_LATE_PAYMENT_zero	average number of days client has late payment (installments)
LANDAREA_MEDI	Median size of landarea
NAME_EDUCATION_TYPE_Lower secondary	client's education type
NAME_EDUCATION_TYPE_Secondary	client's education type
NAME_FAMILY_STATUS_Married	client's family status
NAME_FAMILY_STATUS_Widow_Separated	client's family status
NAME_HOUSING_TYPE_Co-op apartment	Housing type
NAME_HOUSING_TYPE_House / apartment	Housing type
NAME_HOUSING_TYPE_Municipal apartment	Housing type
NAME_HOUSING_TYPE_Office apartment	Housing type
NAME_HOUSING_TYPE_Rented apartment	Housing type
NAME_INCOME_TYPE_Employed	Client's income is employment
OCCUPATION_TYPE_1_Manager	Occupation
OCCUPATION_TYPE_2_Professionals	Occupation
OCCUPATION_TYPE_8_Operators_Assemblers	Occupation
OCCUPATION_TYPE_9_Elementary_Occupation	Occupation
OWN_CAR_AGE_11-15	Car age
OWN_CAR_AGE_16-20	Car age
OWN_CAR_AGE_21-25	Car age
OWN_CAR_AGE_26-30	Car age
POS_CNT_INSTALLMENTS_FUTURE_1-15	Remaing count of installments for POS_CASH loans
POS_CNT_INSTALLMENTS_FUTURE_16-30	Remaing count of installments for POS_CASH loans
POS_CNT_INSTALLMENTS_FUTURE_31-45	Remaing count of installments for POS_CASH loans
POS_CNT_INSTALLMENTS_FUTURE_46-60	Remaing count of installments for POS_CASH loans
POS_CNT_INSTALLMENTS_FUTURE_76+	Remaing count of installments for POS_CASH loans
PREV_AVG_DPD_DAYS	Average DPD days in the previous applications
REGION_RATING_CLIENT_W_CITY_3	Rating of client's location is 3
REG_CITY_NOT_LIVE_CITY_0	Registered address is the same as contact address
YEARS_EMPLOYED_11-15	Years employed
YEARS_EMPLOYED_31+	Years employed
YEARS_EMPLOYED_6-10	Years employed