Data-Driven Credit Risk Modeling: Predicting Probabilities of Default and Assigning Credit Scores

Quynh Lan Nguyen
ID: 501276889

Toronto Metropolitan University

Introduction

- Why credit risk assessment?
- The importance of predictive analytics finance.
- The scope and objectives of the project

Research Questions

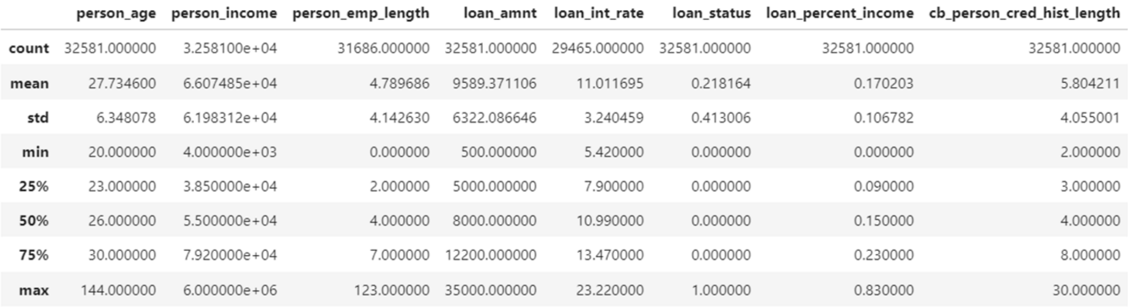
- **Determinants of Default Probability:** What factors are most predictive of loan defaults?

- **Development of an Interpretable Scorecard:** How can we construct a scorecard that transparently assesses credit risk?

- **Model Validation and Reliability:** How dependable is the credit risk model I've developed?

Dataset Overview

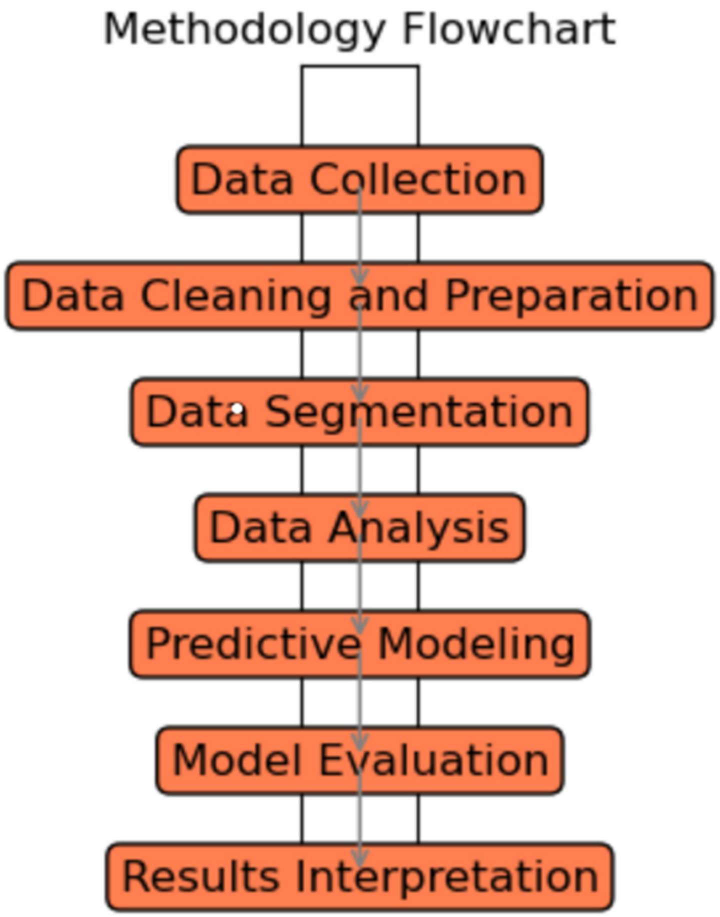
- Over 32,000 consumer loan transactions
- Each transaction is detailed across 12 distinct attributes



	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	32581.000000	3.258100e+04	31686.000000	32581.000000	29465.000000	32581.000000	32581.000000	32581.000000
mean	27.734600	6.607485e+04	4.789686	9589.371106	11.011695	0.218164	0.170203	5.804211
std	6.348078	6.198312e+04	4.142630	6322.086646	3.240459	0.413006	0.106782	4.055001
min	20.000000	4.000000e+03	0.000000	500.000000	5.420000	0.000000	0.000000	2.000000
25%	23.000000	3.850000e+04	2.000000	5000.000000	7.900000	0.000000	0.090000	3.000000
50%	26.000000	5.500000e+04	4.000000	8000.000000	10.990000	0.000000	0.150000	4.000000
75%	30.000000	7.920000e+04	7.000000	12200.000000	13.470000	0.000000	0.230000	8.000000
max	144.000000	6.000000e+06	123.000000	35000.000000	23.220000	1.000000	0.830000	30.000000

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

Methodology



Methodology Flowchart

Data Collection

Data Cleaning and Preparation

Data Segmentation

Data Analysis

Predictive Modeling

Model Evaluation

Results Interpretation

Model Development

- Logistic Regression
- Decision Tree
- Random Forest
- GaussianNB
- Nearest Neighbors
- SVM

Model Evaluation

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.90	0.75	0.78	0.76
Random Forest	0.94	0.97	0.73	0.83
GaussianNB	0.85	0.66	0.60	0.63
Nearest Neighbors	0.89	0.83	0.61	0.70
Logistic Regression	0.87	0.76	0.59	0.66
SVM	0.92	0.92	0.68	0.78

Business Impact

- Being able to foresee and mitigate potential losses before they even materialize
- Financial institutions can confidently expand their lending portfolios, empowering more businesses, and fueling economic growth.
- Transforming data into a decision-making tool.

Limitations, Challenges and Future Research

- Limitation in the scope of the dataset
- Challenges in balancing model complexity with interpretability.
- The need for ongoing research
- The need to integrate alternative data sources.

Conclusions

- Random Forest model helps predict loan defaults.
- Decision Trees model is powerful in developing an interpretable credit risk scorecard, which helps to make well-informed lending decisions.
- Random Forest is the most accuracy and reliable model.

Thank you!

Toronto
Metropolitan
University