Navigating Risk: Insights from Machine Learning Models in Loan Approval for Canadian Banks

Angelia, I.
Longo Faculty of Business
Humber College
Etobicoke, Canada
N01536053@humber.ca

Edward, R.
Longo Faculty of Business
Humber College
Etobicoke, Canada
N01600536@humber.ca

Moromisato, R.
Longo Faculty of Business
Humber College
Etobicoke, Canada
N01578865@humber.ca

***Abstract*— The housing market, a cornerstone of economic stability, faces significant risks amid soaring prices and mounting mortgage delinquencies in Canada. Drawing parallels to the 2008 housing crisis, this paper examines the intersection of risk management and technological innovation in loan approval processes. Considering the exponential risk and complexity of home loans, we proposed a machine learning model to better assess home loan approvals in a more accurate and efficient way.**

Keywords: Loan Prediction, Machine Learning, Logistic Regression, Naive Bayes, CART, Random Forest, KNN.

INTRODUCTION

For many, owning a home represents a significant milestone—a tangible symbol of financial stability and achievement. However, the path to homeownership often involves taking on a mortgage, a financial arrangement that spans decades and involves a partnership between homeowners and financial institutions. This partnership has traditionally been mutually beneficial: homeowners gain access to valuable property while banks profit from interest payments on loans.

Yet, the history of the housing market is rife with turbulence, with the most notable example being the 2008 housing bubble. The collapse of the subprime mortgage market in the United States triggered a global financial crisis, exposing the dangers of reckless lending and speculative excess.

Today, similar warning signs are flashing in the Canadian housing market. Record-low interest rates and lenient lending standards during the pandemic have fueled soaring home prices, raising concerns about the sustainability of the housing boom. As policymakers and experts voice alarm, the specter of another housing bubble looms large, threatening economic stability.

In this uncertain landscape, the role of financial institutions in managing risk is paramount. Banks and lenders must carefully assess the creditworthiness of borrowers to avoid repeating past mistakes. To this end, many are turning to advanced machine learning models to augment traditional risk assessment methods.

This paper explores the intersection of risk management and technological innovation in the context of loan approval. By leveraging the predictive capabilities of machine learning algorithms, financial institutions aim to better identify and mitigate potential sources of risk.

LITERATURE REVIEW

Researchers have expressed growing concerns about Canada's current housing market, noting key distinctions from the 2008 housing bubble in the United States. While speculative activity plays a role, Canada's housing bubble issues primarily stem from a demand-supply imbalance, exacerbated by rapid population growth outpacing housing construction. Notably, low-interest rates offered by banks, especially during the COVID-19 pandemic, have fueled demand. Now, with interest rates on the rise, many Canadians are struggling to keep up with mortgage payments, potentially leading to loan defaults. According to a recent report by Equifax [1], mortgage delinquency rates have surged year-on-year, increasing by 135.2% in Ontario and 62.2% in British Columbia. The trend is particularly pronounced among individuals under the age of 36, especially those who locked in rates during the pandemic and now face mortgage renewals resulting in significant payment hikes ranging from $457 to $680.

A bursting housing bubble can inflict severe economic damage, with the potential to plunge an entire nation into recession. Canada's vulnerability to such a scenario is underscored by its consistent ranking among the top 15 cities with the highest Global Real Estate Index score and Housing-Risk Indicator, as highlighted by both UBS and The Economist [2]. Additionally, leading indicators like the housing-risk indicator have placed Canada at the forefront of concern. Countless studies, including those by [3, 4, 5], have sounded the alarm, warning of the inevitable recession that could follow a housing bubble burst.

Learning from the 2008 housing crisis, Canada faces significant risks from its own housing bubble burst, echoing the catastrophic fallout experienced by financial giants like Lehman Brothers and Bear Sterns. The repercussions would extend far beyond individual homeowners, rippling through banks, lenders, and the broader economy. Plummeting home values and escalating mortgage defaults could inflict staggering losses on financial institutions, potentially triggering credit tightening and economic contraction. The interconnectedness between the housing market and other sectors would amplify the fallout, exacerbating shrinking consumer confidence and spending, while unemployment rates would surge, accompanied by a wave of business bankruptcies. For instance, following the 2008 housing crash, the United States saw a 4.3% GDP decline and doubled unemployment rates, with lingering effects that drained over $2 trillion in global economic growth [6, 7]

In the banking sector, Machine Learning has emerged as a transformative tool for tackling complex tasks that may surpass human capabilities. With the deluge of real-time data streams, Machine Learning seamlessly integrates into banking operations. Danske Bank's adoption of Machine Learning for Fraud Detection stands as a prime example, showcasing a remarkable 60% reduction in false positives—from 1200 to 480—prior to its implementation. Moreover, their accuracy rate has surged from 40% to 50%, underlining the efficacy of Machine Learning in enhancing fraud detection capabilities. Similarly, Postbank leverages Machine Learning to automate loan administration processes, including underwriting and validation tasks. This implementation has drastically reduced human intervention to just 5% of cases, while achieving operational speeds 2.5 times faster than traditional methods, all with minimal error rates [8].

DATASET

Data Collection

The dataset, collected by Konapure [9] for Dream Housing Finance company, consists of a train and test set, each with 13 columns. The columns and encoded values are as follow:

a. Loan_ID
b. Gender - Female: 0; Male: 1
c. Married – No: 0; Yes: 1
d. Dependents – 0: 0 ; 1: 1; 2: 2; 3+:3
e. Education – Graduate: 0; Not Graduate: 1
f. Self_Employed – No: 0; Yes: 1
g. ApplicantIncome
h. CoapplicantIncome
i. LoanAmount
j. Loan_Amount_Term
k. Credit_History: No: 0; Yes: 1
l. Property_Area: Rural: 0; Semiurban: 1; Urban: 2
m. Loan_Status (only train set)

Diagrams

A) Training Set

As part of our analysis, 12 histograms have been generated to explore the training set data from the Konapure housing loan dataset.

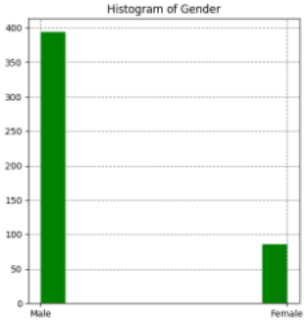


Fig.1: Histogram of Gender

The first histogram displays data regarding the gender of the applicant for home loans from the Konapure dataset. The vertical axis denotes increments of 50 up to 400, while the horizontal axis distinguishes between two groups: “Male” and “Female”. The left green bar representing males reaches a count of approximately 390 applicants. The right green bar representing females reaches a count of approximately 85 applicants. This graphical depiction reveals that there is a notable imbalance between the number of male and female applicants for home loans within the dataset. The data underscores an apparent gender gap, revealing a significant disparity with a higher number of male applicants compared to female applicants.

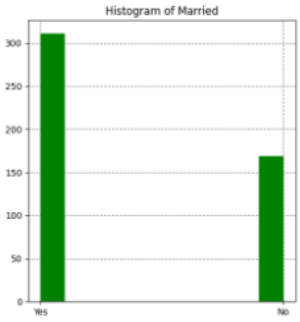


Fig. 2: Histogram of Married

The second histogram illustrates the distribution of married and unmarried applicants in the Konapure dataset. The vertical axis represents increments of 50 up to 350, while the horizontal axis distinguishes between "Yes" and "No" responses indicating marital status. The histogram clearly shows a significantly higher number of married applicants compared to unmarried ones, with approximately 320 "Yes" responses and 170 "No" responses. This disparity highlights a larger proportion of married applicants. Understanding marital status distribution is crucial for analyzing the dataset, as it influences outcomes such as home loan approvals, credit scores, and employment opportunities. Married individuals may be perceived as having a more stable income due to potential dual sources of employment and assets. Strong credit histories for both partners can enhance loan approval chances and terms. However, marital status alone does not determine loan approval.

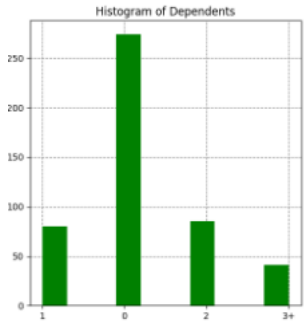


Fig. 3: Histogram of Dependents

The third histogram generated provides visual representation for the distribution of dependents among applicants in the Konapure dataset. The vertical axis denotes increments of 50 up to 300, while the horizontal axis represents the number of dependents, categorized as 1, 0, 2, and 3+. The tallest green bar corresponds to individuals or entities with no dependents. This indicates a higher frequency of such cases within the dataset. The next highest bar represents individuals with two dependents, followed by the

green bars representing applicants with one dependent and three or more dependents.

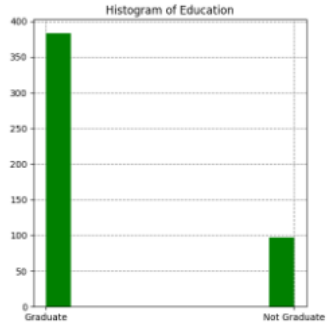


Fig.4: Histogram of Education

The fourth histogram generated provides visual representation for the distribution of graduate and not graduate applicants in the Konapure dataset. The vertical axis denotes increments of 50 up to 400, while the horizontal axis distinguishes between two groups: “Graduate” and “Not Graduate” applicants with respect to their level of education. The number of applicants for home loans with graduate-level education represented by the green bar on the left is significantly higher than the number of applicants for home loans who do not have graduate-level education represented by the green bar on the right. Graduate-level applicants reach a count of approximately 380 while applicants that are not graduates reach a count of approximately 90.

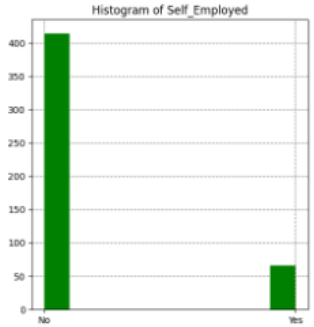


Fig. 5: Histogram of Self Employed

The fifth histogram generated provides visual representation for the distribution of self-employed applicants in the Konapure dataset. The vertical axis denotes increments of 50 up to 450, while the horizontal axis distinguishes between two groups: “No” and “Yes” as to whether the applicant is self-employed. The number of applicants who are not self-employed is notably higher than those that are. The self-employed applicants represented by the green bar on the left reach about 410, while the number of applicants not self-employed reaches about 65.

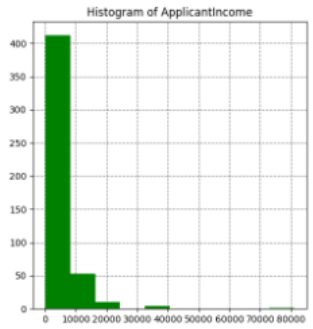


Fig. 6: Histogram of Applicant Income

The sixth histogram generated provides visual representation for the distribution of applicant income in the Konapure dataset. The vertical axis denotes increments of 50 up to 450, while the horizontal axis denotes income by increments of 10000 up to 90000. The histogram shows that applicants with an income level between 0 and 10000 submit the highest number of applications. This is shown by the tallest green bar on the left. Those with income levels greater than 10000 that apply for a home loan are significantly lower.

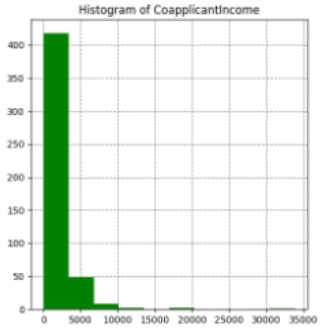


Fig. 7: Histogram of Coapplicant Income

The seventh histogram generated provides visual representation for the distribution of coapplicant income in the Konapure dataset. The vertical axis denotes increments of 50 up to 450, while the horizontal axis denotes income by increments of 5000 up to 35000. The histogram shows that coapplicants with an income level between 0 and 5000 submit the highest number of applications. This is shown by the tallest green bar on the left. Those with income levels greater than 5000 that apply for a home loan are significantly lower. This follows a similar distribution to the primary applicant distribution shown in the previous figure.

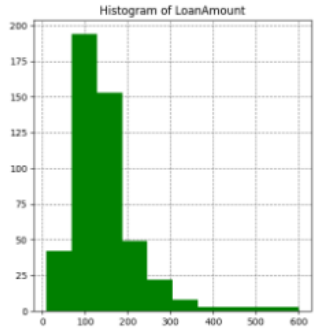


Fig. 8: Histogram of Loan Amount

The eighth histogram generated provides visual representation for the distribution of loan amount in the Konapure dataset. The vertical axis denotes increments of 25 up to 200, while the horizontal axis denotes loan amount by increments of 100 up to 600. The largest number of applications are submitted for loan amounts between 50 to 200. Loan applications for other amounts are significantly lower.

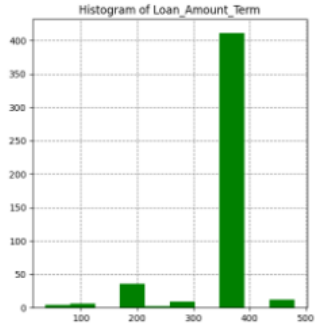


Fig. 9: Histogram of Loan Amount Term

The ninth histogram generated provides visual representation for the distribution of loan terms in the Konapure dataset. The vertical axis denotes increments of 50 up to 400, while the horizontal axis denotes loan amount term by increments of 100 up to 500. Loan terms of 400 carry the largest number of applicants compared to loan terms of 100, 200, 300, and 500. This figure is represented by the tallest green bar.

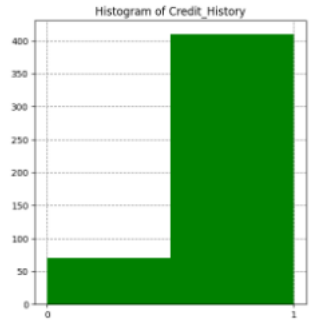


Fig. 10: Histogram of Credit History

The tenth histogram in the Konapure dataset visually represents the distribution of credit history. The vertical axis ranges from 0 to 450 in increments of 50, while the horizontal axis spans credit history values from 0 to 1. Notably, applicants with a credit history are substantially more numerous than those without a credit history.

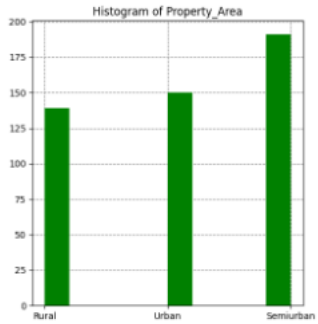


Fig. 11: Histogram of Property Area

The eleventh histogram generated provides visual representation for the distribution of property area in the Konapure dataset. The vertical axis denotes increments of 25 up to 200, while the horizontal axis denotes the area of the property labelled as “Rural”, “Urban”, and “Semiurban”. Most of the applications submitted are for semiurban properties with a count of approximately 190. This is followed by applications for properties located in urban areas and rural areas.

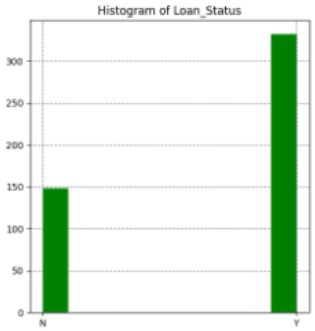


Fig. 12: Histogram of Loan Status

The twelfth histogram generated provides visual representation for the distribution of the loan status in the Konapure dataset. The vertical axis denotes increments of 50 up to 350, while the horizontal axis denotes “N” or “Y” for rejection or approval of the loan application, respectively. The histogram shows that it is highly likely for a home loan application to be approved. The green bar on the right representing the number of approved loan applications is approximately twice as large as the green bar on the left representing the number of loan applications that have been rejected.

B) Loan Status

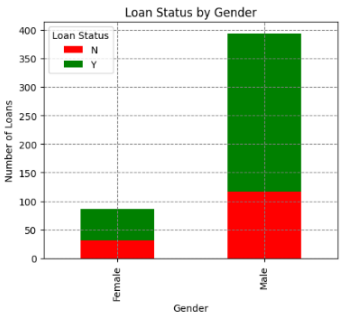


Figure 13: Stacked Bar Chart of Loan Status by Gender

The stacked bar chart above provides a visual representation of the loan status by gender in the Konapure dataset. The vertical axis denotes the number of loans by increments of 50 up to 400, while the horizontal axis denotes gender labelled” Female” and” Male”. The stacked bar chart once again shows that male applicants submit the largest number of home loan applications. Interestingly, the approval rate for male applicants is higher than the approval rate for female applicants. 70.56% of male applicants are approved for home loans compared to 62.79% of female applicants.

Figure 14: Stacked Bar Chart of Loan Status by Marital Status

The stacked bar chart above provides a visual representation of the loan status by marital status in the Konapure dataset. The vertical axis denotes the number of loans by increments of 50 up to 350, while the horizontal axis denotes marital status as" No" for not being married and" Yes" for being married. Like the histogram above, the stacked bar chart shows that the number of loans submitted by applicants who are married are larger than the number of applicants who are not married. Married applicants have a higher approval rate than applicants who are not married. 72.99% of married applicants are approved for home loans compared to 62.13% of unmarried applicants.

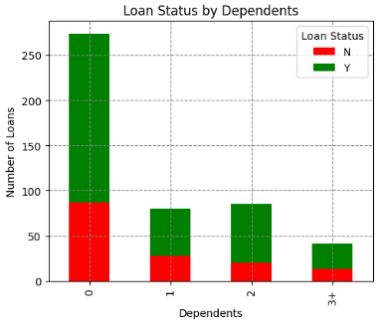


Figure 15: Stacked Bar Chart of Loan Status by Dependents

The stacked bar chart above provides a visual representation of the loan status by number of dependents in the Konapure dataset. The vertical axis denotes the number of loans by increments of 50 up to 300, while the horizontal axis denotes the number of dependents as 0, 1, 2, and 3+. As mentioned earlier, the number of applicants submitted by applicants with 0 dependents is the largest, followed by applicants with 2 dependents, 1 dependent, and 3 or more dependents. Surprisingly, applicants with 2 dependents had the highest approval rate of 76.47% followed by an approval rate of 68.29% for applicants with 3 or more dependents, 68.25% for applicants with no dependents, and 65% for applicants with 1 dependent.

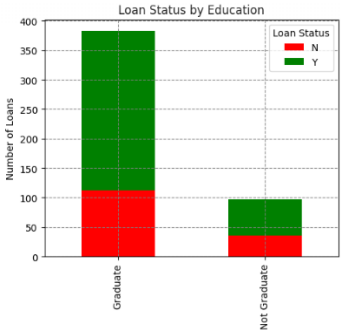


Figure 16: Stacked Bar Chart of Loan Status by Education Level

The stacked bar chart illustrates loan status based on education level in the Konapure dataset. The vertical axis ranges from 0 to 400 in increments of 50, while the horizontal axis distinguishes between "Graduate" and "Not Graduate" categories. Once again, we observe a higher number of home loan applications from graduate applicants compared to non-graduates. Moreover, the approval rate for graduates is higher, standing at 70.76%, in contrast to the 62.89% approval rate for non-graduates. Understanding these trends is vital for assessing the relationship between education level and loan approval.

Figure 17: Stacked Bar Chart of Loan Status by Self-Employed Status

The stacked bar chart illustrates loan status categorized by self-employment status in the Konapure dataset. The vertical axis ranges from 0 to 450 in increments of 50, while the horizontal axis distinguishes between "No" and "Yes" labels for self-employment. It's evident from the chart that there are more loan applications from non-self-employed applicants compared to self-employed ones. Additionally, the approval rate for home loans is higher among non-self-employed applicants, standing at 69.81%, compared to 65.35% for self-employed applicants. These insights shed light on the relationship between self-employment status and loan approval.

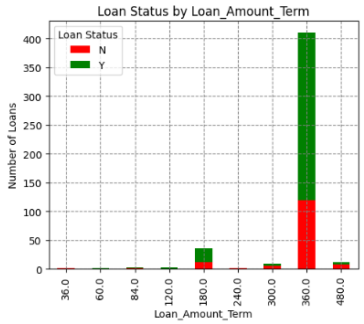


Figure 18: Stacked Bar Chart of Loan Status by Loan Term

The stacked bar chart above provides a visual representation of loan status by loan amount term in the Konapure dataset. The vertical axis denotes the number of loans by increments of 50 up to 450 while the horizontal axis denotes the following loan amount terms: 36, 60, 84, 120, 180, 240, 300, 360, and 480. The chart shows that the number of applicants with loan terms of 360 is the largest with the second largest being loan terms of 180. The loan approval rate is the greatest for terms of 360 at 71.05%, while the loan approval rate for terms of 180 is at 66.67%.

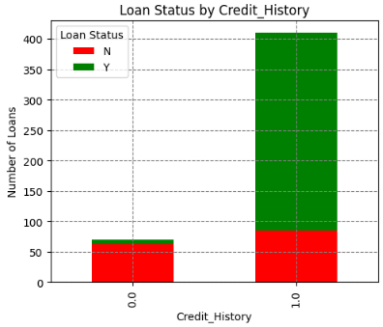


Figure 19: Stacked Bar Chart of Loan Status by Credit History

The stacked bar chart above provides a visual representation of loan status by credit history in the Konapure dataset. The vertical axis denotes the number of loans by increments of 50 up to 450 while the horizontal axis denotes the credit history of 0 and 1. The chart shows that the number of applicants is significantly larger than the number of applicants with no credit history. The loan approval rate is at 79.27% for applications with history of credit while the loan approval rate is only 10% for applications with no history of credit.

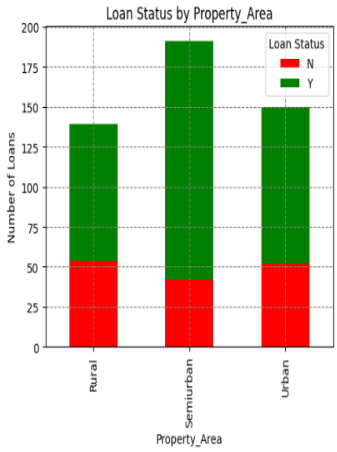


Figure 19: Stacked Bar Chart of Loan Status by Property Area

The stacked bar chart above provides a visual representation of loan status by property area in the Konapure dataset. The vertical axis denotes the number of loans by increments of 25 up to 200 while the horizontal axis denotes the area of the property labelled as “Rural”, “Semiurban”, and “Urban”. The chart shows that the number of applications submitted for home loans against properties in semiurban areas is the largest. The second largest number of applications are submitted for properties in urban areas followed by rural areas. The approval rate follows the same pattern as the number of applications submitted with semiurban properties having a home loan approval rate of 78.01%, urban properties having an approval rate of 65.33%, and rural properties having an approval rate of 61.15%.

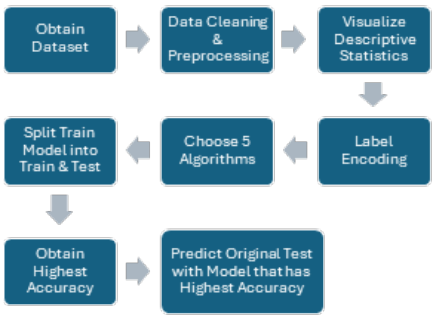
METHODOLOGY

Data Cleaning

Initially, the train set contained 614 rows, reduced to 480 rows after cleaning. The test set had 367 initial rows, reduced to 289 rows after cleaning.

Data processing for model

To maintain consistency across all the models, all columns were used as predictors except for loan_id and a split was done for the training set with a ratio of 80:20 and random_state of 10. The train set is now divided into another train and test data, with 384 and 96 data values respectively. 5 algorithms were used in this research, and the model with the highest accuracy would be used to predict the original test dataset that does not have the loan_status value.



Flowchart 1: Project Flowchart

Algorithm Used

A) Classification and Regression Trees (CART)

CART is a model that operates like a decision tree, dividing data into branches based on the most optimal split points. These splits are determined by GINI impurity, a measure of how often a randomly chosen element would be incorrectly classified. By choosing splits that minimize GINI impurity, CART constructs a tree that effectively segments the data into homogeneous groups, enhancing predictive accuracy.

B) Random Forest Classification

Random Forest is a machine learning model that shares similarities with CART, utilizing the concept of decision trees that make splits at optimal points. However, Random Forest distinguishes itself as a collection of multiple decision trees, each trained on a slightly different subset of the same dataset. During training, each tree is constructed using a random subset of features at each split, which helps to reduce overfitting and promote diversity among the trees. When making predictions, the model aggregates the outputs of these individual trees, typically through a majority vote for classification tasks or averaging for regression tasks. This ensemble approach often yields superior generalization performance compared to a single decision tree.

C) K-Nearest Neighbour (KNN)

KNN is a data-driven machine learning model that makes no assumptions about the underlying data. It calculates the Euclidean distance between each predictor variable of the test data point and the training data points. By sorting distances from smallest to largest, KNN identifies the nearest neighbors. Classification of the test value is then determined by the majority class among the nearest K neighbors.

D) Naïve Bayes

Naïve Bayes, a probabilistic classifier, treats each predictor independently, assigning probabilities to each class based on the presence of individual features. Despite its simplifying assumption of feature independence, Naïve Bayes effectively categorizes data by calculating the likelihood of each class given the input features, making it particularly efficient for large datasets.

E) Logistic Regression

Logistic Regression adopts the idea of linear regression, except the outcome would be a binary of 0 or 1 where it can be used for both explanatory and predictive tasks. Unlike Naïve Bayes where every predictor is assumed to be independent of each other, logistic regression considers the relationship between one independent variable and another. The Logistic Regression Model are built with the following formula:

$$p = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}} \quad (1)$$

With this formula, a probability between 0 and 1 is first calculated. Afterwards, the odds of Y = 1 is then calculated with the following formula:

$$Odds(Y = 1) = \frac{1}{1 - p} \quad (2)$$

A class would then be assigned based on the cutoff value to determine which class the value belongs to.

RESULTS

Upon reviewing the analysis, we find that both logistic regression and random forest exhibit the highest accuracies among all models, as demonstrated below:

```
Logistic Regression Accuracy: 0.8021
Multinomial Naive Bayes Accuracy: 0.5833
CART Accuracy: 0.7812
Random Forest Accuracy: 0.8021
KNN Accuracy: 0.7917
```

Figure 20: Accuracies for all models

Additionally, upon examining the confusion matrices for all models, a consistent pattern emerges overall, the models perform well but tend to exhibit a positive class bias. This is evidenced by the occurrence of more false positives (Type I errors) than false negatives. Such findings suggest a potential risk associated with approving loans for non-qualified individuals.

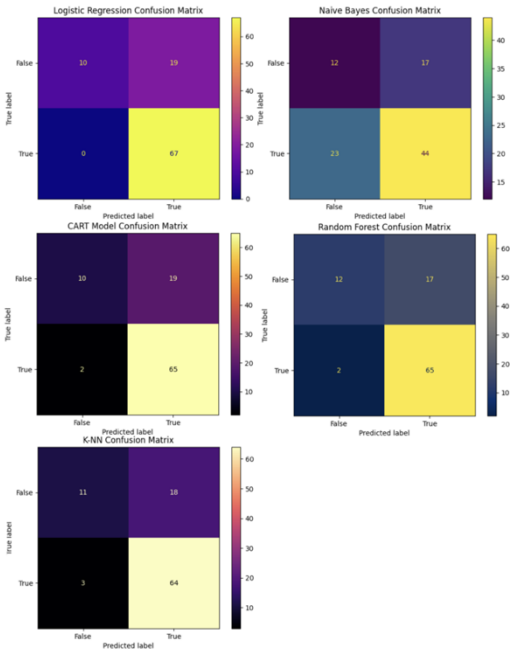


Figure 21: Confusion Matrices for all models

Lastly, we deploy the top-performing model Random Forest, in this instance to predict loan approval status in the test set:

```
['Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y'
 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y'
 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'N' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'N'
 'N' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'N' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y'
 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y'
 'Y' 'Y' 'Y' 'N' 'N' 'N' 'Y' 'Y' 'N' 'N' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y'
 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y'
 'Y' 'N' 'Y' 'N' 'Y' 'Y' 'Y' 'N' 'N' 'Y' 'N' 'Y' 'N' 'Y' 'N' 'N' 'Y' 'Y'
 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'N' 'Y' 'Y' 'N' 'Y' 'N' 'Y' 'Y' 'Y'
 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y'
 'N' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'N' 'N' 'Y' 'Y' 'N' 'Y' 'N' 'Y'
 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'N'
 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y'
 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'N' 'N' 'Y' 'Y' 'Y'
 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N' 'Y' 'Y' 'N' 'N'
 'Y' 'Y' 'N' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'Y' 'N'
 'Y']
```

Figure 22: Predicted values on the test set

Utilizing this model enables financial institutions to enhance loan approval assessments, leading to improved efficiency through more accurate and expedited decision-making. This, in turn, helps to reduce costs and streamline processes for optimized operations.

ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Predicted_Loan_Status
5720	0	110.0	360.0	1.0	2	Y
3076	1500	126.0	360.0	1.0	2	Y
5000	1800	208.0	360.0	1.0	2	Y
3276	0	78.0	360.0	1.0	2	Y
2165	3422	152.0	360.0	1.0	2	Y
...	...	...	...	...	...	...
2269	2167	99.0	360.0	1.0	1	Y
4009	1777	113.0	360.0	1.0	2	Y
4158	709	115.0	360.0	1.0	2	Y
5000	2393	158.0	360.0	1.0	0	N
9200	0	98.0	180.0	1.0	0	Y

Figure 23: Test set with predicted values

CONCLUSION

In conclusion, the utilization of machine learning for automating bank loan approval processes holds promise in enhancing accuracy and mitigating risks associated with loan defaults, particularly when dealing with complex predictor combinations that may be challenging to assess manually. Despite achieving an 80% accuracy rate in our train model, it's essential to acknowledge the limitations posed by the short timeframe and the relatively small dataset available for analysis. Additionally, variations in predictor variables across different banks may impact the generalizability of our findings. Nonetheless, we remain optimistic that the insights gained from this study could offer valuable assistance to banks in Canada, empowering them to make loan approval decisions with greater precision and efficiency.

REFERENCES

[1] S. Rosa, "Canadian mortgage-holders increasingly missed payment in Q4, Equifax says", CTV News, 2024. [Online]. Available: https://www.ctvnews.ca/business/canadian-mortgage-holders-increasingly-missed-payments-in-q4-equifax-says-1.6794802

[2] Mortgage Sandbox, "Is the canadian real estate market a bubble? Here are the risks to consider", 2024. [Online]. Available: https://www.mortgagesandbox.com/risk-in-the-canadian-real-estate-market

[3] B. LaCerda, A. Singh, S. Mintah, and G. Pinel, "Canada housing market outlook: more struggles ahead", Moody's Analytics., 2023. [Online]. Available: https://www.moodysanalytics.com/whitepapers/pa/2023/rps-ma-canada-housing-market-outlook-more-struggles-ahead

[4] I. Poshnjari, "A housing bubble burst would be worse in Canada than U.S.: Rosenberg", BNN Bloomberg, 2022. [Online]. Available: https://www.bnnbloomberg.ca/a-housing-bubble-burst-would-be-worse-in-canada-than-u-s-rosenberg-1.1841896

[5] G. Suhanic, "Posthaste: the coming recession will be a tale of housing versus commodities", Financial Post, 2024. [Online]. Available: https://financialpost.com/news/canada-recession-about-housing-versus-commodities

[6] J. Weinberg, "The great recession and its aftermath", Federal Reserve History, 2013. [Online]. Available: https://www.federalreservehistory.org/essays/great-recession-and-its-aftermath#:~:text=Effects%20on%20the%20Broader%20Economy,-The%20housing%20sector&text=The%20decline%20in%20overall%20economic,recession%20since%20World%20War%20II

[7] R. Merle, "A guide to the financial crisis – 10 years later", The Washington Post, 2018. [Online]. Available: https://www.washingtonpost.com/business/economy/a-guide-to-the-financial-crisis--10-years-later/2018/09/10/114b76ba-af10-11e8-a20b-5f4f84429666_story.html

[8] Projectpro, "15 projects on machine learning applications in finance", 2024. [Online]. Available: https://www.projectpro.io/article/projects-on-machine-learning-applications-in-finance/510

[9] R. Konapure, "Home loan approval", Kaggle, n.d. [Online].Available: https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval