

2024

THE FRONT LINE RESEARCH REPORT

We aim to refine payment predictions, benefiting financial risk management strategies.

PREPARED BY: THE FRONT LINE
(KATE ANGLIN, JACOB BAKER, AND NICK WATTS)

Table of Contents

Table of Contents	1
Overview	2
Data Overview	3
Logistic Regression	4
Multinomial Naïve Bayes	5
Decision Tree	6
Random Forest	7
Neural Networks	8
Top 3 Models	9
Testing	10
Analysis Overview	11
Overall Conclusion	12
Thank You!	

Overview

The Front Line has been tasked with predicting which clients are most likely to default on their loans.

Our project revolves around participating in the Home Credit - Credit Risk Model Stability competition, which aims to predict which clients are more likely to default on their loans. By leveraging machine learning, we seek to provide consumer finance providers with a reliable method for assessing default risk. Ultimately, this will enable more informed lending decisions.

Overview

Using our machine learning configurations, we intend to accurately identify individuals at a higher risk of defaulting on a loan. Leveraging statistical and machine learning methods, we aim to develop a Python source that can predict loan risk effectively. However, the challenge lies in ensuring the stability of these models over time, as clients' behaviors constantly evolve.

Summary

Our project aims to provide innovative solutions to predicting loan default risk while recognizing the importance of stability and reliability in the financial industry's decision-making processes.

Project Limitations:

Despite our efforts, it is essential to acknowledge certain limitations inherent in the project:

1. **Data Availability:** Limited availability of comprehensive credit data for specific client segments may hinder the accuracy of our predictions. Access to crucial information, such as credit history or financial stability indicators is necessary for our models to avoid challenges in accurately assessing default risk.
2. **External Factors:** Our predictions may be influenced by external factors beyond our control such as macroeconomic conditions, regulatory changes, or unforeseen events. These external factors could introduce uncertainty into our predictions and impact their reliability in real-world applications.

Data Overview

We were supplied with 26.77 gigabytes of data. We chose to train with one csv file that contains personal information for over 1.5 million loans.

Overview

The dataset analyzed for this project totals 26.77 gigabytes. It contains 33 training files, but for simplicity and computational feasibility, we chose one file with over 1.5 million loan instances. That file is called `train_person_1.csv`. There were 1,526,660 data instances in this CSV file. There were 37 features; the first feature is the `case_id`. The `case_id` values correspond to the `case_id` values in the `train_base` CSV file. The `train_base` CSV file contains a few features, but the most important are the `case_id` and `target`. Intuitively, the `case_ids` from `train_person_1` correspond to the `case_ids` in `train_base`. For example, if a loan instance from `train_person_1` has a `case_id` value of 1337, the data instance in `train_base` with a `case_id` value of 1337 is the same person. Finally, the `target` is the `y` and last feature in `train_base`. The `target` feature will have a value of 0 for a person who has not defaulted and a value of 1 for a person who has defaulted. Class 1 and 0 have respective counts of 47,994 and 1,478,664, making up 3.14% and 96.86% of the dataset.

Summary

We have the personal information and default status from a lot of borrowers. Many of these were given as strings, so one-hot encoding was required. We have a clear `X` and `y` here, so this is perfect for artificial intelligence.

Project Limitations:

The dataset underwent stratified pruning, reducing the number of instances to 15,268 for training, validation, and testing purposes. This process maintained the original distribution of default and non-default instances. An 80:20 split was employed for training and validation, resulting in 12,828 instances for development (`dev.csv`) and 2,957 instances for testing (`test.csv`). The reduced feature set, downsized from 33 to 16, aims to mitigate runtime issues while retaining crucial information for predictive analysis. Since one-hot encoding was used on this dataset, our computers could not handle the computations without pruning. The pruned datasets still provide a balanced yet concise representation for subsequent model development and evaluation.

Logistic Regression

Information & Data:

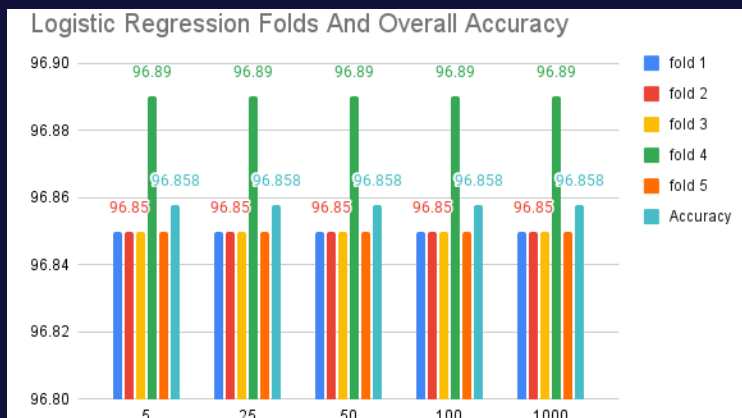
Logistic regression is pivotal in predicting loan default probabilities for the Home Credit - Credit Risk Model Stability competition, leveraging the dataset to train and test models. The dataset's characteristics are crucial in shaping logistic regression's performance in classifying loan applicants' default probabilities.

Configurations:

Changes: Iterations for logistic regression were adjusted from 5 to 1000 to assess performance impact. Despite stable performance, no significant algorithm alterations were made; instead, the focus shifted to fine-tuning parameters like iteration count.

What happened: Logistic regression consistently performed at around 96.85% accuracy across iterations, suggesting convergence to an optimal solution. Additional iterations did not notably improve predictive performance.

Interpreted: The consistent accuracy across iteration levels indicates that logistic regression likely reached an optimal solution, emphasizing its reliability in predicting loan default probabilities.



Findings:

The analysis highlights logistic regression's robustness in predicting loan default probabilities, with consistent accuracy observed across varying iteration levels. This stability indicates that logistic regression has effectively captured the underlying patterns in the data relevant to loan default prediction.

Analysis:

Graphical representations of logistic regression accuracy across different iteration levels provide insights into the algorithm's behavior and performance trends. The stability in accuracy underscores logistic regression's suitability for modeling loan default risks, offering finance providers a reliable tool for risk assessment.

Conclusion:

Logistic regression emerges as a reliable and stable approach for predicting loan default probabilities within the Home Credit - Credit Risk Model Stability competition context. Its consistent performance, interpretability, and computational efficiency position logistic regression as a valuable tool for finance providers in managing credit risk and making informed lending decisions.

Multinomial Naïve Bayes

Information & Data:

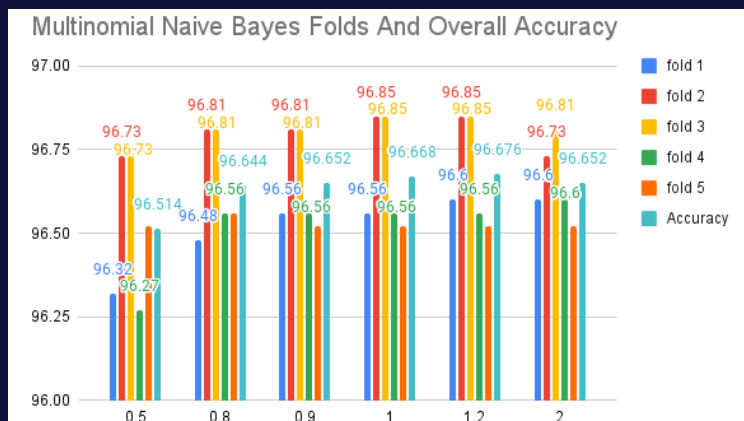
Multinomial Naive Bayes, an alternative algorithm utilized in predicting loan default probabilities for the Home Credit - Credit Risk Model Stability competition, employs a probabilistic classification approach. Like logistic regression, it relies on the dataset for training and testing. The features and characteristics of the dataset significantly impact the model's effectiveness in categorizing loan applicants' default probabilities.

Configurations:

Changes: The alpha parameter for Multinomial Naive Bayes was adjusted from 0.5 to 2, incrementing by 0.1, to assess its impact on model performance. Values were selected based on their influence on accuracy, aiming to find the optimal value for improved performance.

What happened: Multinomial Naive Bayes demonstrated varying accuracy levels across tested alpha values. Accuracy initially increased steadily from 0.5 to 1.2 but declined after reaching an alpha of 1.2. The final alpha chosen was 2, yielding an accuracy of approximately 96.652%, which was a significant drop from 1.2's accuracy of 96.676.

Interpreted: The model's accuracy pattern with changing alpha values indicates sensitivity to the smoothing parameter. Through iterative testing, an optimal alpha of 2 was identified, balancing between underfitting and overfitting. This signifies peak performance before experiencing a decline, reflecting a trade-off between bias and variance in the model.



Findings:

The analysis of Multinomial Naive Bayes reveals the sensitivity of model performance to changes in the alpha parameter. Optimal performance is achieved when balancing bias and variance, with the selected alpha value of 2 indicating a trade-off point before experiencing diminishing returns.

Analysis:

Graphical representations of Multinomial Naïve Bayes accuracy across different alpha values provide insights into the algorithm's behavior and performance trends. The observed pattern underscores the importance of parameter tuning in optimizing model performance and avoiding overfitting.

Conclusion:

Multinomial Naïve Bayes utilizes a probabilistic approach to predict loan default probabilities, effectively leveraging dataset features. The model optimizes performance through systematic alpha parameter adjustments, enhancing risk assessment accuracy for finance providers. Despite its effectiveness, Multinomial Naïve Bayes demonstrated slightly lower accuracy than logistic regression, prompting exploration of alternative algorithms like Decision Trees to enhance further predictive performance in the Home Credit - Credit Risk Model Stability competition.

Decision Tree

Information:

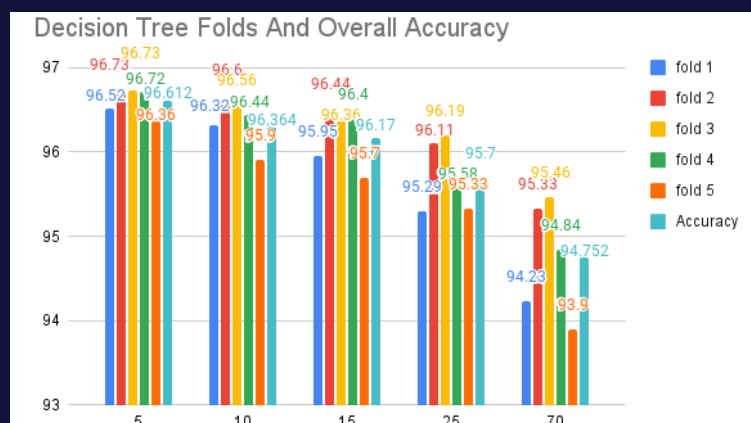
Decision tree models are utilized to predict loan default probabilities for the Home Credit - Credit Risk Model Stability competition, offering a rule-based classification approach similar to multinomial naive Bayes and logistic regression. These models rely on dataset features for training and testing, with the dataset's characteristics significantly impacting the model's ability to classify loan applicants' default probabilities.

Configurations:

Changes: The depth parameter of the Decision Tree was adjusted to assess its impact on model performance. Depths ranged from a high initial value down to lower values, exploring various levels of tree complexity to determine the optimal depth for improved accuracy.

What happened: Across different depths tested, the Decision Tree exhibited varying levels of accuracy. Surprisingly, as the depth decreased, accuracy improved, with the highest accuracy achieved at lower depths. This indicates that a more straightforward decision tree structure resulted in better generalization performance, likely avoiding overfitting issues associated with deeper trees.

Interpreted: The observed accuracy trend suggests that reducing the tree's depth led to a more generalized and robust model, improving its ability to classify loan default probabilities accurately. By simplifying the decision tree structure, the model avoided capturing noise and irrelevant patterns in the training data, enhancing performance on unseen data.



Findings:

The analysis of Decision Tree models highlights the importance of balancing model complexity with generalization performance. While deeper trees may capture more complex patterns, simplifying the tree structure can improve overall model performance by reducing the risk of overfitting.

Analysis:

Graphical representations of Decision Tree accuracy across different depth levels underscore the inverse relationship between model complexity and generalization performance. The observed improvement in accuracy with decreasing depth emphasizes the need for careful parameter tuning to optimize model performance.

Conclusion:

Starting with a high depth, Decision Tree models showed improved performance as depth decreased, indicating that shallower structures may better predict loan default probabilities, avoiding overfitting. Further exploration of alternatives like Random Forest, employing ensemble methods to mitigate overfitting, may enhance predictive performance in the Home Credit - Credit Risk Model Stability competition.

Random Forest

Information & Data:

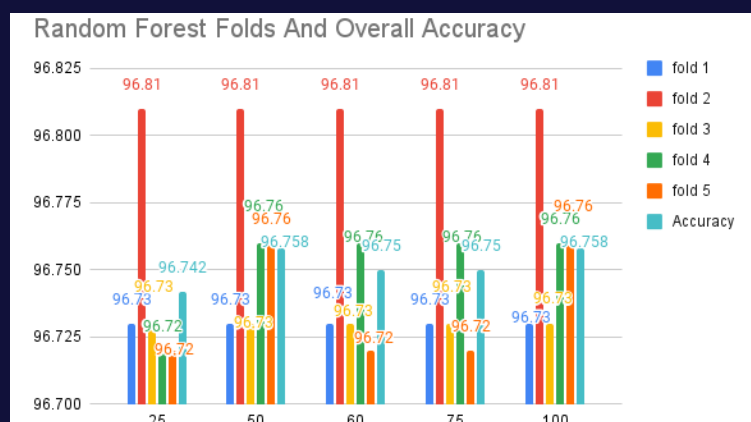
Random Forest models are employed to predict loan default probabilities for the Home Credit - Credit Risk Model Stability competition, utilizing an ensemble method combining multiple decision trees. Like other algorithms, such as Decision Trees, Random Forest relies on dataset features for training and testing, with dataset characteristics significantly influencing model performance.

Configurations:

Changes: The number of trees in the Random Forest was adjusted to assess its impact on model performance, varying from 25 to 100 trees. This exploration aimed to determine the optimal number of trees for improved accuracy in predicting loan default probabilities.

What happened: Across different numbers of trees tested, the Random Forest exhibited consistently high accuracy levels. Notably, accuracy remained consistently high at approximately 96.75% to 96.76%, with minor fluctuations observed. The highest accuracy was achieved at 50 and 100 trees, indicating robust performance at these values.

Interpreted: The observed accuracy trend suggests that Random Forest models maintain stable and high performance across varying numbers of trees. The peak accuracy at 50 and 100 trees demonstrates the effectiveness of ensemble methods in capturing diverse patterns within the dataset. This consistency indicates the reliability of Random Forest in predicting loan default probabilities.



Findings:

The Random Forest models achieved the highest accuracy compared to other algorithms explored thus far, with peak performance observed at 50 and 100 trees. This consistency highlights the robustness of Random Forests in handling complex classification tasks, such as predicting loan default probabilities.

Analysis:

Graphical representations of Random Forest accuracy across different numbers of trees reaffirm the stability and high performance of the model. The consistent accuracy levels underscore the effectiveness of ensemble methods in leveraging diverse decision trees to improve predictive performance.

Conclusion:

Given the superior performance of Random Forest models, particularly at 50 and 100 trees, further exploration into alternative algorithms like Neural Networks is warranted. Neural Networks, with their complex architecture and ability to capture nonlinear relationships, may provide intense competition in predictions, especially considering similarities in their construction with Random Forest algorithms. Exploring these alternatives can enhance predictive performance and provide robust solutions for risk assessment in the Home Credit - Credit Risk Model Stability competition.

Neural Networks

Information:

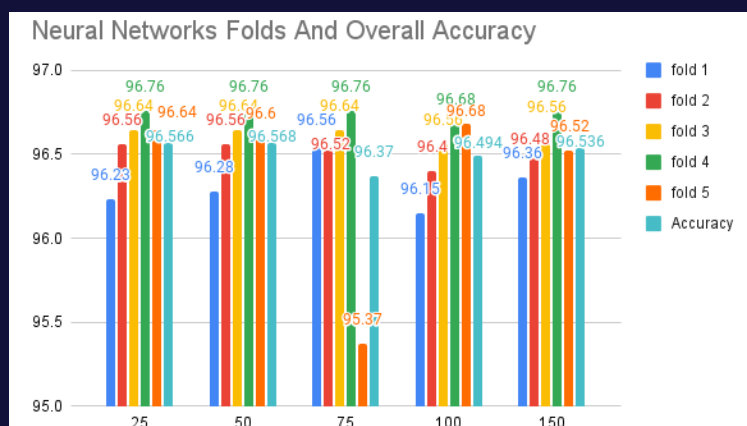
Neural Network models are utilized to predict loan default probabilities for the Home Credit - Credit Risk Model Stability competition, employing a complex architecture designed to capture nonlinear relationships within the dataset. Like other algorithms, Neural Networks rely on dataset features for training and testing, with dataset characteristics significantly influencing model performance.

Configurations:

Changes: The number of hidden layers in the Neural Network was adjusted to assess its impact on model performance, varying from 25 to 150 hidden layers. This exploration aimed to determine the optimal number of hidden layers for improved accuracy in predicting loan default probabilities.

What happened: Across different numbers of hidden layers tested, the Neural Network exhibited varying levels of accuracy. However, the accuracy achieved could have been higher than other algorithms explored in the project. Fluctuations in accuracy were observed, with no clear trend indicating significant improvement with increasing hidden layers.

Interpreted: The observed accuracy trend suggests that the Neural Network models did not perform as well as expected in predicting loan default probabilities for this project. Despite variations in the number of hidden layers, the accuracy remained below the benchmark set by other algorithms. This indicates that Neural Networks may not be the optimal choice for this particular classification task.



Findings:

The Neural Network models achieved lower accuracy than other algorithms explored in the project, indicating their limited effectiveness in predicting loan default probabilities. The inability to surpass the benchmark set by other algorithms suggests that Neural Networks may not be suitable for this project.

Analysis:

Graphical representations of Neural Network accuracy across different numbers of hidden layers reveal fluctuations in performance without a clear improvement trend. The inconsistent accuracy levels underscore the challenges in optimizing Neural Network models for this classification task.

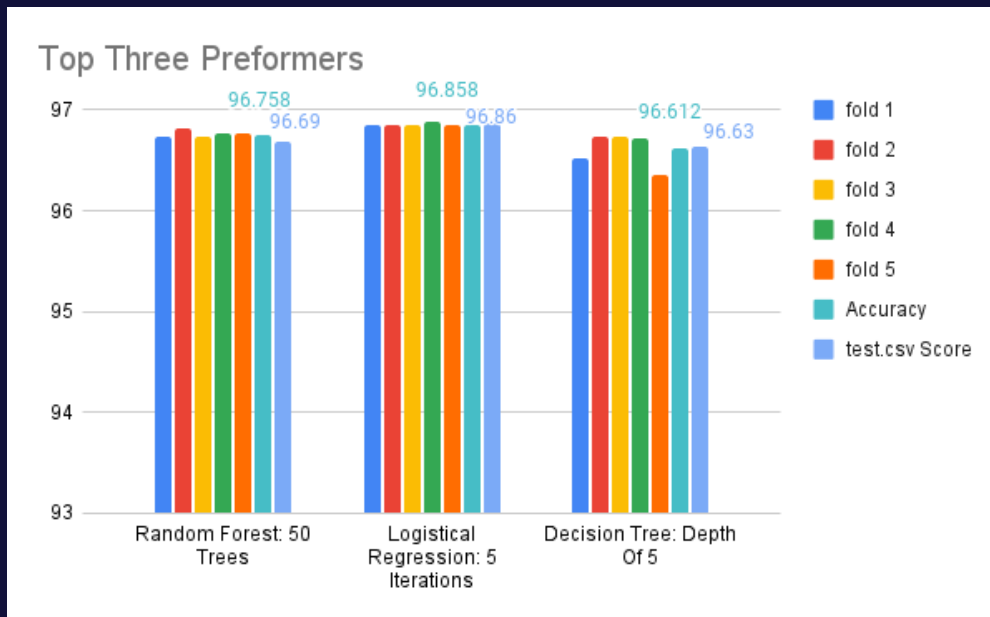
Conclusion:

As the last algorithm tested, Neural Network models' performance in predicting loan default probabilities fell short compared to other explored algorithms. Despite some improvement, their accuracy remained below the benchmark set by top-performing models. Therefore, we have decided to revert to the top three algorithms overall: Logistic Regression, Decision Tree, and Random Forest, due to their superior performance and stability. By optimizing these top-performing algorithms further, we aim to enhance predictive performance and provide robust risk assessment solutions for the competition.

Top 3 Models

Information:

The top three performing algorithms in predicting loan default probabilities for the Home Credit competition are Logistic Regression (5 Iterations), Random Forest (50 Trees), and Decision Tree (Depth of 5). Logistic Regression achieved the highest accuracy of 96.86%, followed closely by Random Forest with 96.69% accuracy and Decision Tree with 96.63% accuracy. Logistic Regression, a linear model, offers simplicity and interpretability, making it suitable for this task. Random Forest, an ensemble method, offers stability against overfitting and handles high-dimensional data effectively. Despite its simplicity, the decision tree offers insights into the importance of features but requires careful tuning to prevent overfitting. The choice of algorithm depends on specific requirements and constraints, with each offering different trade-offs in accuracy, interpretability, and computational efficiency.



Second

First

Third

The Best would be Logistic Regression at 5 iterations with an accuracy of 96.858

The Second would be Random Forest at 50 trees with an accuracy of 96.758

The Third would be Decision Tree at depth of 5 with an accuracy of 96.612

Conclusion:

Although Logistic Regression emerged as the highest-performing model in terms of accuracy, Random Forest and Decision Tree models also showed promise in predicting loan default probabilities. The selection of the most suitable algorithm depends on factors such as interpretability, computational efficiency, and the trade-off between bias and variance. Moving forward, efforts will focus on optimizing Logistic Regression to enhance its predictive performance further while exploring the potential of ensemble methods like Random Forest for improved stability and generalization.

Testing

Information:

The three top-performing algorithms in predicting loan default probabilities for the Home Credit competition are Logistic Regression (5 Iterations), Random Forest (50 Trees), and Decision Tree (Depth of 5). When submitting to Kaggle, the code broke; however, the algorithms worked ideally outside the Kaggle environment.

Decision Tree (Depth of 5):

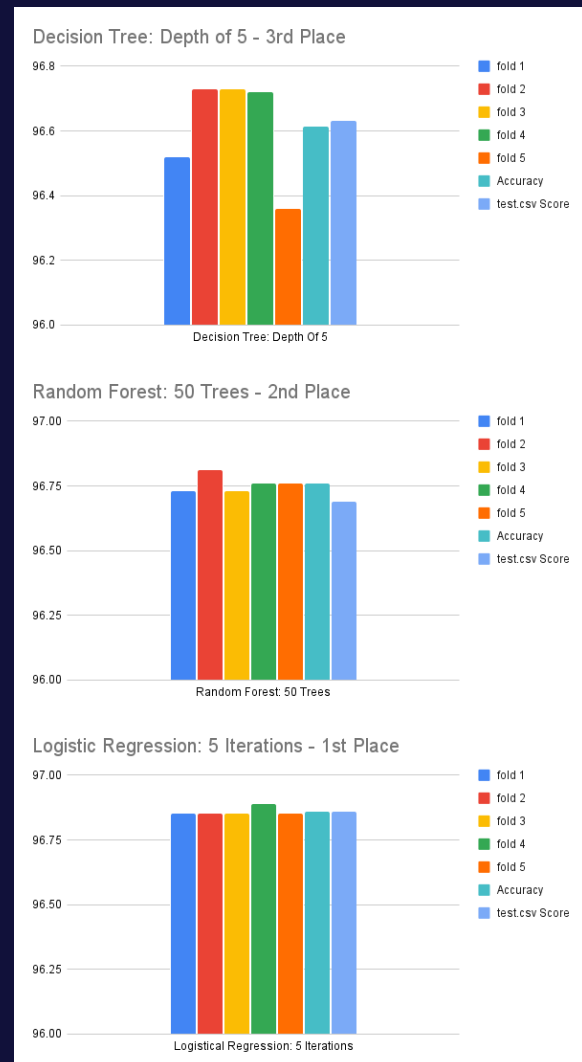
- Achieving an accuracy of 96.63%, this model had a test.csv score of 96.612.

Random Forest (50 Trees):

- This model obtained an accuracy of 96.69%, obtaining a test.csv score of 96.758.

Logistic Regression (5 Iterations):

- With an accuracy of 96.86%, this model scored 96.858 on the test.csv dataset.



Conclusion:

While Logistic Regression emerged as the best-performing model in terms of accuracy, Random Forest and Decision Tree also demonstrated strong performance. However, formatting issues were encountered during the submission of predictions to Kaggle. This problem is suspected to arise from modifying files outside the Python environment, emphasizing the importance of careful data handling within the code for future submissions.

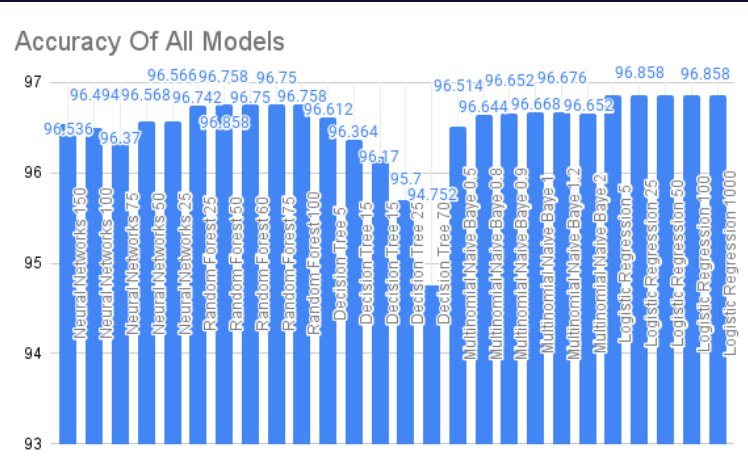
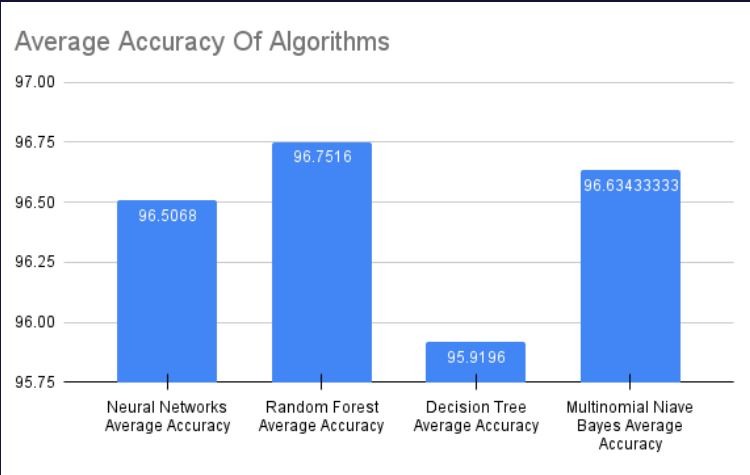
Analysis Overview

Information:

This project employs machine learning techniques to predict credit debit likelihood of non-payment, aiming to contribute to risk management in financial institutions. The dataset, meticulously prepared and split into development and test sets, contains crucial information for modeling credit debit payment behavior. Various algorithms and hyperparameters are systematically evaluated through cross-validation to enhance predictive accuracy, ultimately developing robust models capable of accurately forecasting credit debit payment outcomes.

Output Data Overview:

The project involved training and testing 26 models to predict loan default probabilities for the Home Credit competition. Logistic Regression emerged as the top-performing model with an average accuracy of 96.858% and a standard deviation of 0.040. Random Forest and Decision Tree also showed promising results, with average accuracies of 96.7516% and 95.9196%, respectively. Multinomial Naive Bayes and Neural Networks exhibited lower average accuracies of 96.6343% and 96.5068%, respectively, with higher standard deviations indicating more significant variability in performance.



Conclusion:

The analysis highlights the effectiveness of Logistic Regression, Random Forest, and Decision Tree algorithms in predicting loan default probabilities. However, formatting issues were encountered while submitting predictions to Kaggle, emphasizing the importance of careful data handling within the code for future submissions. Moving forward, efforts will further optimize the top-performing algorithms to enhance predictive performance and provide robust risk assessment solutions for financial institutions.

Overall Conclusion

Summary of all Information:

Logistic Regression stood out as the top-performing model, boasting an average accuracy of 96.858% with a slight deviation of 0.040. Random Forest and Decision Tree also performed well, scoring average accuracies of 96.7516% and 95.9196%, respectively. However, Multinomial Naive Bayes and Neural Networks demonstrated slightly lower average accuracies of 96.6343% and 96.5068%, respectively, with higher standard deviations indicating more significant variability in performance.



Best Algorithm:

Logistic Regression (5 iterations) emerged as the top-performing algorithm, achieving an accuracy of 96.86%. Its simplicity and interpretability make it a reliable tool for predicting loan default probabilities. Despite its straightforward nature, Logistic Regression outperformed more complex models with remarkable accuracy.

Problems and Errors:

- Formatting issues were encountered while submitting predictions to Kaggle, highlighting the importance of careful data handling within the code for future submissions.


Ideas to Continue:

#	Recommendations	Priority
1	Further optimization of the top-performing algorithms to enhance predictive performance.	HIGH
2	Exploration of alternative algorithms or techniques to address formatting issues and improve model stability.	HIGH
3	Apply methods to prevent overfitting and improve generalization.	MEDIUM
4	Fine-tune models with extensive parameter searches.	LOW
5	Enhance models with additional features or transformations.	MEDIUM



Thank you!

Thank you for reviewing our project report, the Credit Risk Model Stability Report. We appreciate your interest and hope our findings provide valuable insights into credit risk management.



Created By: The Front Line
(Kate Anglin, Jacob Baker, & Nick Watts)