Yihao Luo, Yiwen Sun, Yiqing Xin
ORIE 5741 - Final Project
Professor He
5/12/2023

Analyzing Loan Application Trends in New York State: Machine Learning Insights into Who Can Get a Loan

Abstract

This paper employs machine learning techniques, such as logistic regression, decision trees, and random forest, to investigate the loan application trends in New York State in 2021. It aims to identify the applicant characteristics that are favored by lenders and have a higher probability of application acceptance. The data used in this study was obtained from the FFIEC Home Mortgage Disclosure Act (HMDA), which contains 787,436 observations with 99 variables of information on borrowers, lenders, demographics, loan details, etc. The paper outlines the data features used, their engineering methods, and the data cleaning and preparation techniques applied. Ultimately, the project aims to assist loan applicants in determining their likelihood of application acceptance, which can benefit both applicants and financial institutions.

1. Introduction

This project aims to analyze loan application trends in New York State in 2021, using machine learning techniques to gain insights into which applicant characteristics are favored by lenders and have a higher probability of application acceptance. Usually, algorithms focus on avoiding high-risk applicants for financial institutions. When borrowers fail to repay loans or meet contractual obligations, the interruption of cash flow and increased cost of loan collection are some problems that arise from credit risk. In this project, however, by identifying the applicant characteristics favored by lenders, the project aims to help applicants determine whether their loan application will be accepted, which can benefit both loan applicants and financial institutions.

2. Data

2.1 Data Introduction

We utilized data from the FFIEC Home Mortgage Disclosure Act (HMDA), an official United States government website that provides information about loan applications and decisions. The dataset we used focuses on New York State in 2021, which contains 787,436 observations. It contains 99 variables of information about borrowers, demographics, loan details, etc.

The response is obtained from the variable called *denial_reason_1*, which records denial reasons for denied applications and approved projects which are denoted as "Not Applicable" in *denial_reason_1*. We created the response variable by assigning 1 to observations that are "Not Applicable" in *denial_reason_1* and 0 to the rest of the observations. In this way, 1 denotes applications that are approved, and 0 denotes denied applications.

Among the remaining 98 variables, we selected essential variables as features based on project purpose, type of variables, and nature of data. In the end, we were able to narrow down to 23 variables that can be classified into six categories: demographics, borrower basics, property nature, the purpose of the loan, contract details, and others (See Table 1. below).

The demographic category contains variables of county code, ethnicity, race, sex, and applicant age. County code records the state-county FIPS code that identifies counties where each loan application took place. The borrower basics categories have variables including income, debt-to-income ratio, conforming loan limit, and purchaser type. The conforming loan limit indicates whether the reported loan amount exceeds the GSE (government-sponsored enterprise) conforming loan limit. Purchaser type records ten types of entities purchasing a covered loan from the institution. The category of property nature records information about properties that people apply for loans, which includes total units, property value, and the loan-to-value ratio. Variables in the purpose of loan category are loan product type, loan purpose, business or commercial purpose. Loan product type records four covered loan or application types and lien status for each observation. Business or commercial purpose shows whether the covered loan or application is primarily for a business or commercial purpose. The contract details category includes specific programs or rules in application processes or contracts, such as preapproval program, reverse mortgage, open-end line of credit, loan amount, interest rate, rate speed, and loan term variables. Preapproval is a preliminary loan commitment from a lender that helps potential home buyers determine their budget and gives them an edge in a competitive market. An open-end line of credit is a type of loan that provides flexible borrowing and repayment options. Other variables that do not belong to specific categories include submission of application and lei. Submission of an application indicates if the applicant or borrower submitted the application directly to the financial institution, and lei records the financial institution's legal entity identifier.

Table 1. Data Feature Table

Column Variable Name	Type	Type Feature Engineering
county_code	Demographic	One hot encoding
derived_ethnicity	Demographic	One hot encoding
derived_race	Demographic	One hot encoding
derived_sex	Demographic	One hot encoding
applicant_age	Demographic	One hot encoding

income	Borrower Basics	Numeric variable
debt_to_income_ratio	Borrower Basics	One hot encoding
conforming_loan_limit	Borrower Basics	One hot encoding
purchaser_type	Borrower Basics	One hot encoding
derived_loan_product_type	Purpose of Loan	One hot encoding
loan_purpose	Purpose of Loan	One hot encoding
business_or_commercial_purpose	Purpose of Loan	One hot encoding
total_units	Property Nature	One hot encoding
property_value	Property Nature	Numeric variable
loan_to_value_ratio	Property Nature	Numeric variable
preapproval	Contract Details	One hot encoding
open-end_line_of_credit	Contract Details	One hot encoding
loan_amount	Contract Details	Numeric variable
interest_rate	Contract Details	Numeric variable
rate_spread	Contract Details	Numeric variable
loan_term	Contract Details	Numeric variable
Submission of Application	Other	One hot encoding
lei	Other	Numeric variable

2.2 Data Cleaning & Preparation

Now that we have all features for our data, we would like to clean and prepare data before analyzing data and building models. In regard to feature engineering, we kept numerical variables in their original values and applied one-hot encoding to categorical variables. In this case, missing values from one-hot coding will be automatically filled.

For numerical variables, we replaced all missing values with corresponding mean values. The reason is that dropping missing values will only keep observations having a response equal to 1, which is sensible since applicants with more complete information are more likely to get loans.

In addition, we added a dummy variable to indicate whether an observation has missing values or not. After data cleaning, the dataset still contains more than 7,000,000 observations with 178 predictors. The number of data points is much larger than the number of features. In this way, we do not face the problem of the curse of dimensionality and can apply all these selected features in our models.

2.3 Exploratory Data Analysis

Before diving into the modeling part, we analyzed the data by exploring its composition and relationships between variables. This can help us identify any trend or pattern in the data and inform the application of further data cleaning and selection of suitable models.

First, we plotted the distribution of the response variable (see Figure 1. below). The histogram shows the percentage of acceptance and rejections in the dataset. 87% of applications are approved, and 13% of applications received rejection. In this way, we have much more data points for approved applications than rejected applications.

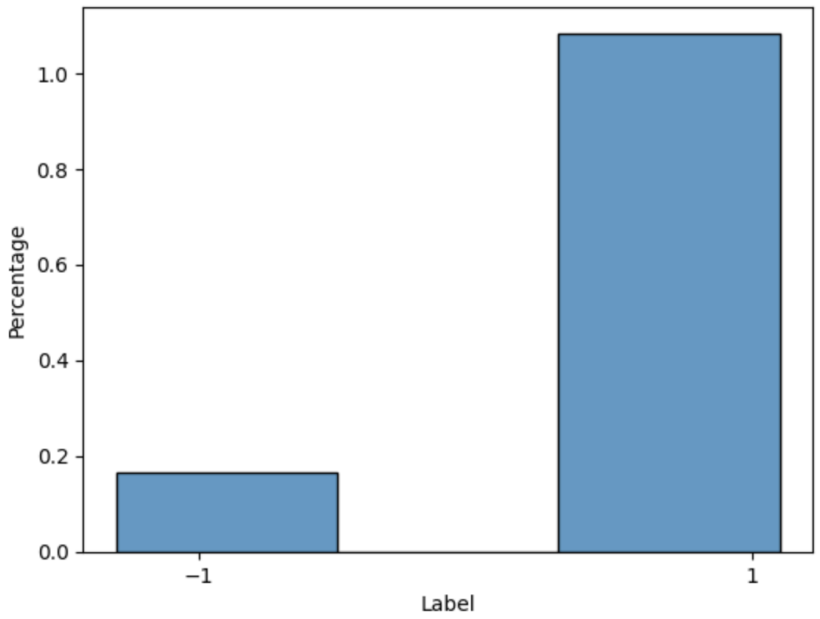

Figure 1. Distribution of Response Variable

In the next step, we created a heatmap of the correlation matrix of the 23 essential variables that were selected from all features (see Figure 2. below). In this graph, off-diagonal values correlation coefficients between each variable and other variables. As the color of a grid gets darker, it means that the correlation between two variables is higher. Most grids have correlation coefficients of less than 0.15 in absolute values, which suggests that they barely correlate with each other. However, some variables have strong correlations, as big as 1, and should be

removed to eliminate the problem of multicollinearity. These variables are an open-end line of credit, submission of application, and business or commercial purposes.

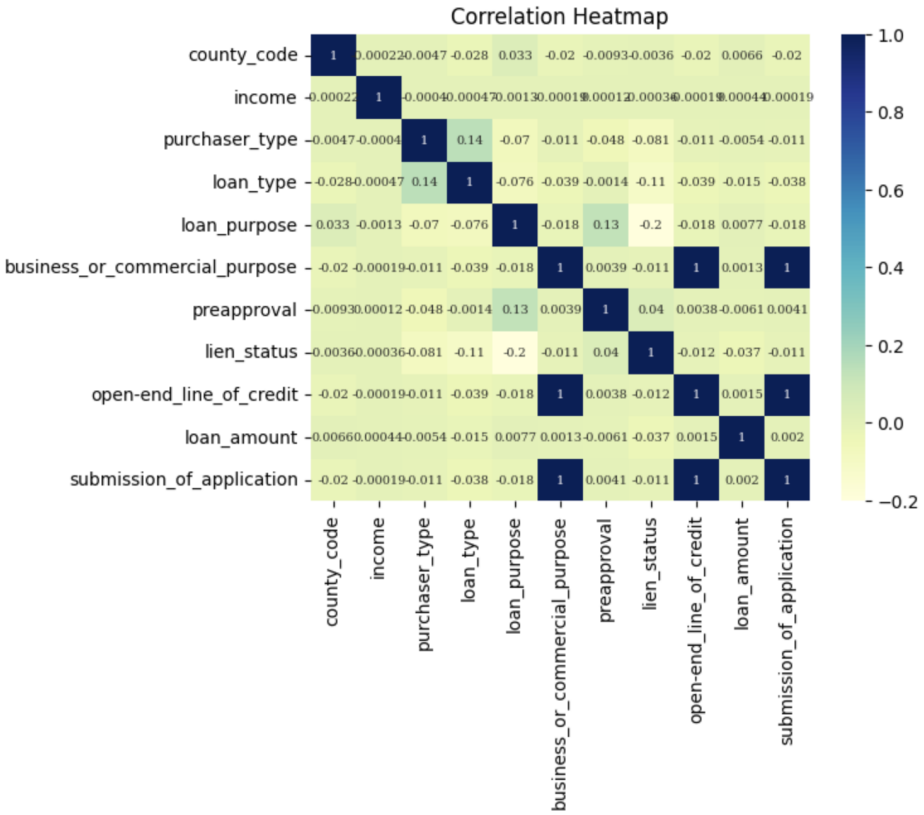

Figure 2. Correlation Heatmap of the 23 Variables

3. Model

The problem aims to determine whether an application for a loan will be accepted or not, so we tried three models that are suitable for this classification problem: logistic regression, decision tree, and random forest tree. Before moving on to each model, we first split cleaned data into training and test sets, with 80% of data in the training set and 20% in the test set.

3.1 Logistic Regression

Logistic regression is a statistical method used to analyze the relationship between a binary dependent variable and one or more independent variables. It estimates the probability of an event occurring by applying a sigmoid function to map real values of a linear function into another value between 0 and 1. It is a type of regression method analysis that is commonly used when the outcome variable takes on only two possible values, such as 1 or -1, yes or no, etc. In this project, the outcome variable is whether a loan application was approved or rejected, and we encoded it into 1 and -1, which makes logistic regression suitable for our analysis.

Then we fitted logistic regression to the training set and made predictions on the test set using the fitted model. To measure the performance of the logistic regression model, we used three measurements: precision, recall, and AUC. Precision is the ratio of the true positive predictions to all the positive predictions made by the model. It measures the model's accuracy in predicting positive cases. Recall refers to the proportion of true positives among all the positive cases in the data, which shows the model's ability to correctly identify positive cases. The receiver operating characteristics (ROC) curve visualizes the tradeoff between the true positive rate and the false positive rate for different threshold values. The area under the ROC curve (AUC), ranging from 0 to 1, measures the model's overall ability to distinguish between positive and negative cases, with high values indicating better performance.

After we made predictions on the training set and compared them with test responses, the model has a precision of 0.83 and a precision of 0.99, which is considered good performance. However, AUC is a little small, only around 0.58 (See Figure 3. below). An AUC of 0.5 suggests that the model performs no better than random guessing, which indicates that the logistic regression model has a poor ability to distinguish between positive and negative cases.

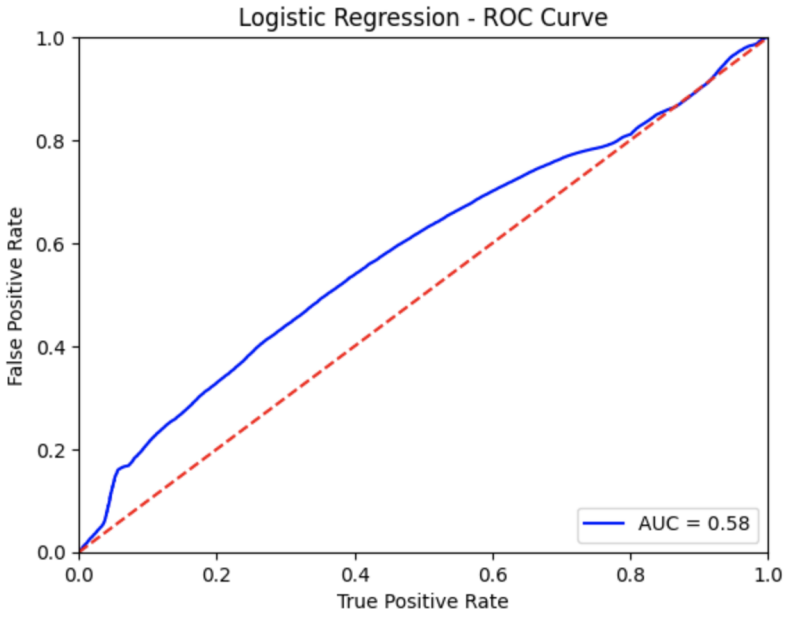

Figure 3. ROC Curve and AUC Score of Logistic Regression

3.2 Decision Tree

The decision tree is a simple yet powerful machine-learning method that can be applied to both classification and regression tasks, depending on the form of the loss function. In this case, we have used the entropy loss to construct a classification tree.

There are several advantages to using decision trees. Firstly, they allow us to easily identify the most relevant features by calculating the average entropy loss on each feature. This is particularly valuable in our case, as our goal is to identify the features that are likely to result in loan rejections and help borrowers improve their financial profiles. Additionally, decision trees have the benefit of making predictions in O(log(n)) time complexity, which makes them suitable for online learning scenarios. Potential borrowers can perform self-assessments and receive instant feedback using decision tree-based models. Furthermore, decision trees can be combined to build powerful learning tools such as random forests and gradient-boosting trees. These ensemble methods aggregate the predictions of multiple trees in different ways, resulting in improved performance and robustness. This makes tree-based models well-suited for our specific task. In summary, decision trees offer valuable features for our task, including feature relevance identification, quick prediction times, and the potential for building more advanced ensemble models.

To avoid overfitting, we aim to find the simplest decision tree that achieves the highest validation accuracy. In order to determine the optimal tree depth, we fit decision trees with different maximum depths and analyze their train and validation loss. The train and validation loss are plotted on the following graph (See Figure 4. below)

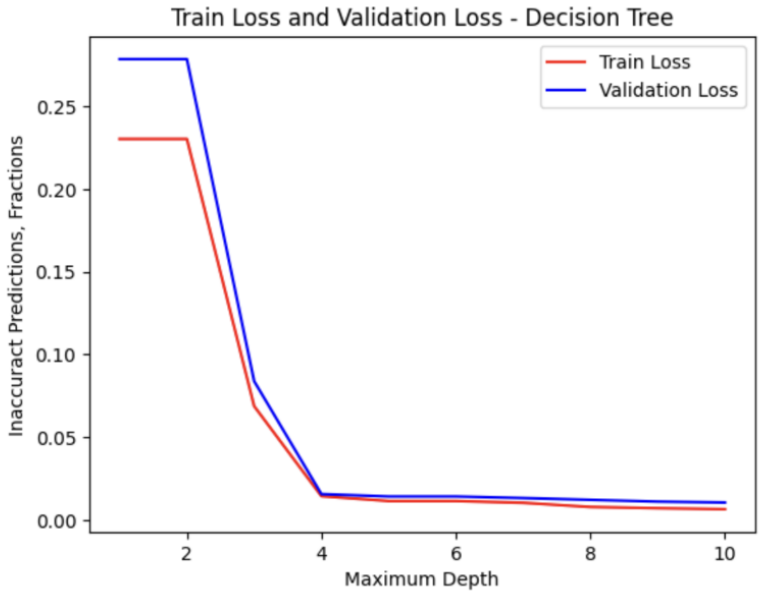

Figure 4. Train and Validation Loss from Decision Tree

Based on the plot, it is evident that a small decision tree with a maximum depth of 4 can achieve a high validation accuracy of 99.2%. Further splits do not result in an increase in validation accuracy. Therefore, we select a maximum depth of 4 for our decision tree (See Figure 5. below)

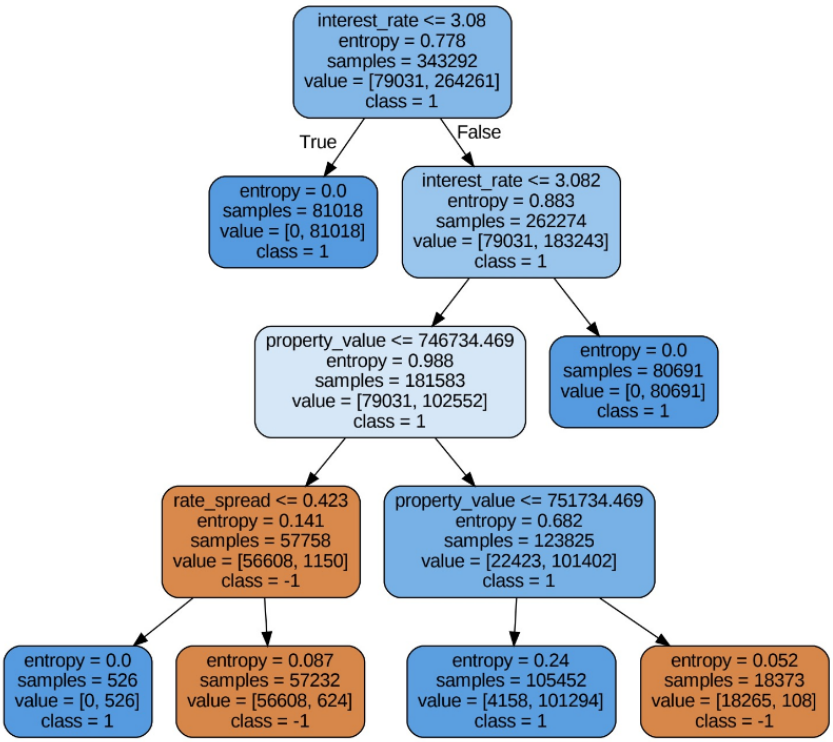

Figure 5. Decision Tree with Maximum Depth of 4

The analysis reveals that interest_rate and property_value are the two most influential variables, contributing to 37.2% and 61.4% of the entropy loss, respectively. These findings align with our expectations: lenders offer lower interest rates to borrowers with lower default risks, resulting in guaranteed loan approval when the interest_rate is below 3.08%. Additionally, higher property values indicate greater borrower wealth and a higher likelihood of meeting financial obligations, increasing the chances of securing a mortgage loan. Consequently, this decision tree highlights the limited number of key variables involved in the loan decision process. To explore other significant factors, we have implemented Random Forest, which will be discussed in the upcoming section.

3.3 Random Forest

While decision trees offer accurate predictions, it's crucial to address their tendency to overfit. To enhance generalizability, reduce model variance, and encourage consideration of other important factors, we have implemented a Random Forest for the same task. The key concept behind Random Forest is to alleviate overfitting by reducing prediction variance. This is achieved by subsampling both data and features from the original dataset to construct multiple trees. Subsequently, predictions are made using these individual trees, and the most frequent prediction is selected as the final prediction. By not utilizing every feature for each decision tree, the model is implicitly compelled to take into account factors beyond just interest_rate and property_value.

As the Training Loss of Random Forest provides an unbiased estimation of the validation loss, we have selected the number of trees based on the training loss. The relationship between the training loss and the number of trees is given in the following plot (See Figure 6. below).

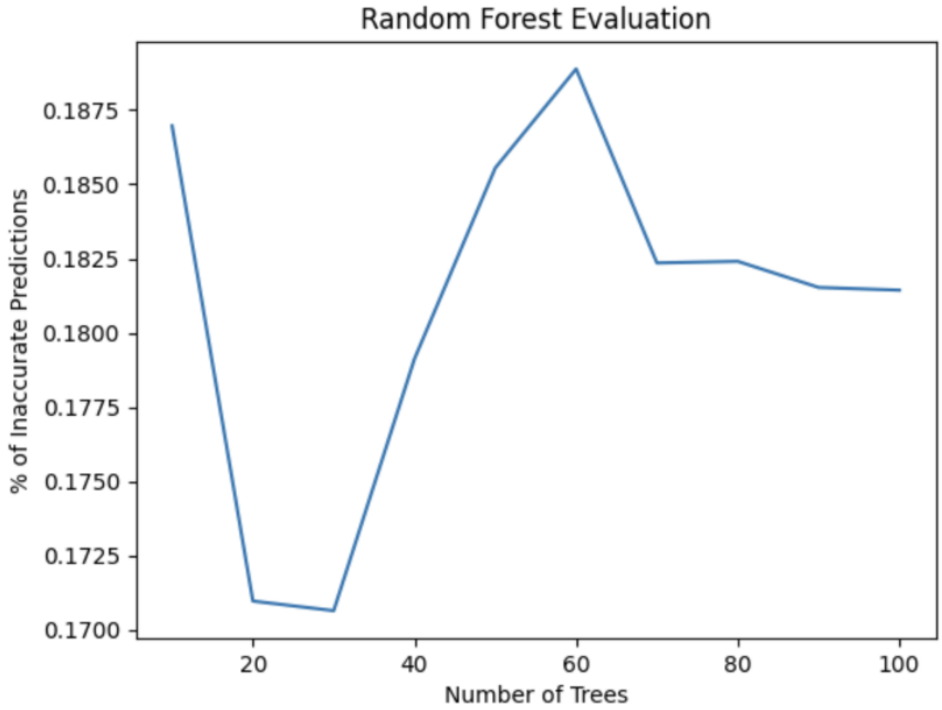

Figure 6. Training Loss VS the Number of Trees From Random Forest Model

It is clear that Random Forest has worse accuracy than individual Decision Tree because it is forced to consider more factors. However, we could still find that selecting the forest size of 30 gives the lowest validation (training) accuracy. By calculating the mean loss of entropy, we summarize the feature importance in the following table (See Table 2. below).

Score: 0.8208755228784824

	feature	importance
3	interest_rate	0.224368
23	debt_to_income_ratio->60%	0.181814
4	rate_spread	0.180363
0	property_value	0.102915
1	loan_to_value_ratio	0.062061
7	debt_to_income_ratio-50%-60%	0.047234
6	income	0.043117
140	open-end_line_of_credit-1	0.016489
141	open-end_line_of_credit-2	0.016173
54	applicant_age-8888	0.011249

Table 2. Feature Importance Table

Although ‘interest_rate’ and ‘property_value’ remain among the most important features, we have identified other important features such as debt_to_income_ratio, loan_to_value_ratio, income, rate_spread, and applicant_age.

Those findings are also intuitive: Applicants who have a higher debt-to-income ratio have more existing financial obligations relative to their income, so the financial institutes might be reluctant to offer them new loans. A higher loan-to-value ratio implies that the home buyer is buying their homes on greater leverage, making them more vulnerable to fluctuations in real estate prices or interest rates, which will become a concern for the financial institutes. When a home buyer is closer to retirement, the financial institution might start to concern about the lender's ability to fulfill long-term financial obligations as they are expected to experience an income drop after retirement and hence become reluctant to offer them new loans.

Overall, Random Forest helps us to identify the key factors that lead to loan rejection. We could leverage those findings to help the borrowers to improve their loan application profile and eventually enable them to get a loan.

4. Conclusion

Based on the above three models involved in this project, the random forest model is a better fit for predicting loan approvals as it considers various factors that lenders use to assess applicants' creditworthiness. To be specific, a random forest model with 30 trees improved the accuracy to 82% by taking into account other factors that influence loan approval, which is a relatively high accuracy.

4.1 Risks & Next Step

Even though the accuracy of the random forest model reached 82%, it still has a probability of misclassifying loan applications. One side of misclassification will result in lending money to high-risk borrowers who are likely to default on their loans, which could lead to financial losses for the lending institution and negatively impact their reputation. On the other hand, rejecting low-risk loan applications could result in missing business opportunities and even negatively affect the institution's reputation. It's important to continuously monitor the model's performance and adjust it based on new information derived from applicants to minimize the risk of misclassification. Additionally, it's important to ensure that the model is not discriminating against certain groups of people based on characteristics such as race or gender.

Since the model is ready, the next step would be to deploy the random forest model on a web platform or integrate it into an existing loan application system to provide instant feedback to applicants on their application's acceptance probability. This will help potential borrowers understand the chances of their loan applications being approved and make informed decisions about their financial future. Moreover, financial institutions can improve their customer services and application process efficiency if they apply the model to their systems.

References

Home Mortgage Disclosure Act. (2021). FFIEC Census Reports Data. Retrieved May 12, 2023, from https://ffiec.cfpb.gov/data-browser/data/2021?category=states.