ORIE 5741 Project Final Report
Credit Approval Prediction

Ruifan Chen
rc654

Xinyue Liu
xl533

Yijun Zhu
yz2745

Abstract

It is of great significance for financial institutions to make wise decisions about which individuals or businesses to offer credit to avoid bad debt. The main dataset is the home mortgage disclosure action provided by HMDA Platform combined with the dataset of median housing prices in counties in NY state. This project applies big data techniques to build machine learning models to make better predictions on whether to offer credit to a client. We develop logistic regression, random forest models, together with Pearson correlation and Lasso feature selection approach for modeling process. Moreover, we further consider feature interaction terms and PCA dimension reduction methods to improve model performance. In the end, the project will conduct a fairness analysis and discuss corresponding limitations and improvements

1 Exploratory Data Analysis

1.1 Data Characteristics

The main dataset for our project is from the HMDA platform, including our label, home mortgage disclosure action, with 98 explanatory variables from 2018 to 2020 in New York State, which has 1825608 entries in total. Among all the 99 variables, there are 86 categorical variables with 25 ordinal data and 61 nominal data, 13 numerical variables including 2 discrete variables, and 11 continuous variables. These variables focus on the three aspects:

1. Financial and demographic background of applicants such as gross annual income, credit score type, ethnicity.

2. Demographic information of the county or tract of applicants including tract-population, median-housing-units.

3. Information about the loan such as loan term, loan amount, and interest rate.

The second dataset we apply is the median house price data of each county in New York state from 2018 to 2020 from the official website of New York state(https://www.ny.gov/). We merge the two datasets to get a combined one with 100 variables in total.

In addition, we plan to consider the influence of the Covid-19 Pandemic on loan approval cases, so we add one nominal variable: "Covid". It is set to be 1 if the application is in 2020, otherwise 0. In this case, we obtain our final dataset, composed of the label "action-taken" and in total 100 explanatory variables from 2018 to 2020 in New York State.

Considering the aim of the project, the current dataset is messy and large that many variables contain high volume null values. Thus, in the first step, we would discuss the quality of the dataset and clean data for our further data analysis.

1.2 Data Cleaning

1.2.1 Missing Values

According to Figure 1.1, there are more than 30 variables which have a proportion of over 50 percent of NA values. For most variables, "NA" values do not have actual meaning because applicants leave them blank when filling in the application form. Moreover, for many of these variables, the information is repetitively inferred in other variables.

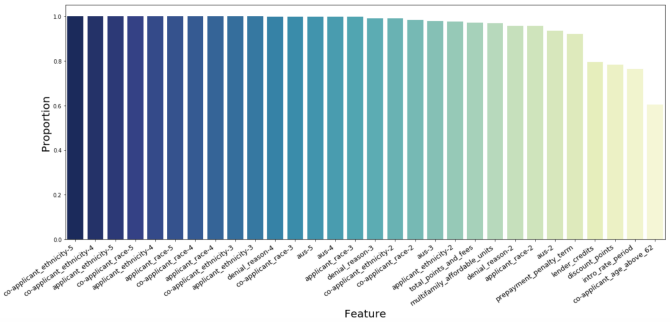

Figure 1.1: Missing values proportion for top 30 variables

To deal with the null values, variables over 50 percent NA values (31 in total) are dropped. For variables containing less than 5 percent of NA values, we delete all the corresponding entries with NA values in the corresponding variable column. Lastly, for those variables with 5 to 50 percent NA values, we fill in the NA values using our estimate by random forest models separately.

1.2.2 Outlier Detection

Outliers in the numerical variables may exert a negative influence on our data visualization and analysis. Thus, we apply the Inter-quartile Range Approach and regard values outside the range of [Lower quartile - 1.5 * Inter-quartile range, Upper quartile + 1.5 * Inter-quartile range] as outliers, and then delete their corresponding entries.

1.2.3 Data Corruption

To facilitate our further analysis and increase the interpretability of our models, we relabel our target variable

"action-taken". Originally, this nominal variable has in total 7 values including Application in progress, Application denied, Application withdrawn, Application approved but not accepted, and so on. In this case, we combine "Application approved but not accepted" and "Purchased loan" as "Approve", combine "Application denied" and "Preapproval request denied" as "Deny", and drop other values' entries. As a result, we obtain a binary label for our target variable.

1.3 Data Visualization

1.3.1 Correlation Analysis

After we have done the basic data cleaning process, we are able to conduct data visualization to explore relationship among variables, and understand the significance of different explanatory variables in credit approval prediction.
First of all, we split variables into categorical variables and numerical variables. For numerical ones, we conduct a correlation heat-map to explore the relationship between each variable and the label. As we can see from Figure 1.2, the label "Action-taken" is relatively highly correlated with features "loan-term" and "median-housing-price".

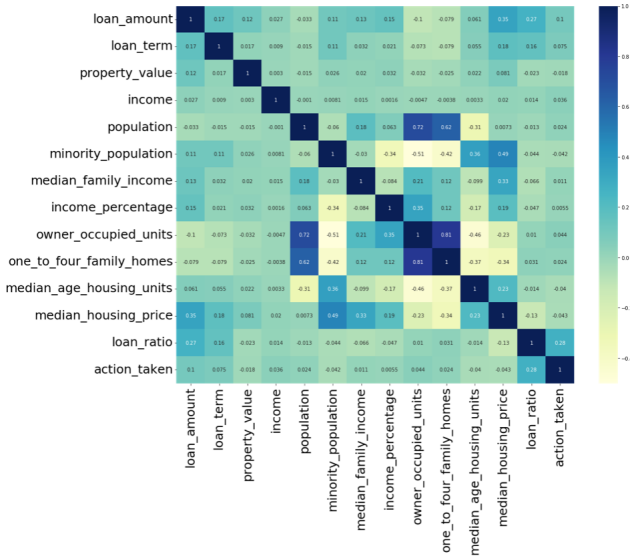

Figure 1.2: Numerical Variable Correlation Analysis

1.3.2 Important Numerical Variable Visualization

According to our findings in Figure 1.2, we consider the two most correlated numerical variables: "loan-term" and "median-housing-price". We create several groups based on value, and plot group bar charts to see their influences on the credit approval rate.
As we can see from Figure 1.3 and Figure 1.4, there is an obvious positive relationship between "loan-term" and "action-taken", and there exists a negative relationship between "median-housing-price" and "action-taken".

1.3.3 Important Categorical Variable Visualization

Intuitively, we choose two important categorical variables to draw bar charts below to see the influences of their different levels on the credit approval rate.

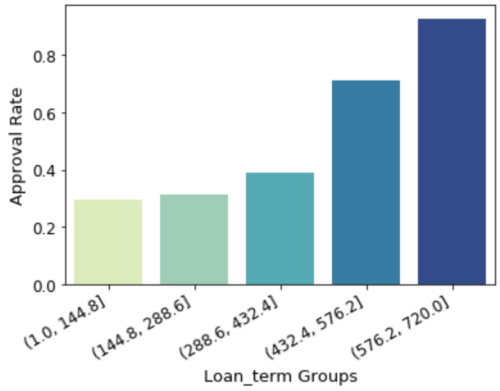

Figure 1.3: Loan-term Variable Group Analysis

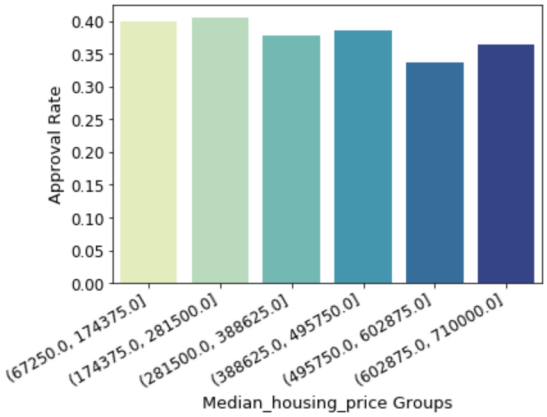

Figure 1.4: Median-housing-price Variable Group Analysis

According to Figure 1.5 and Figure 1.6, we could find that as the value varies, "credit score" and "sex" have a great impact on our target variable, and we would conduct further analysis on their relationship in the following sections.

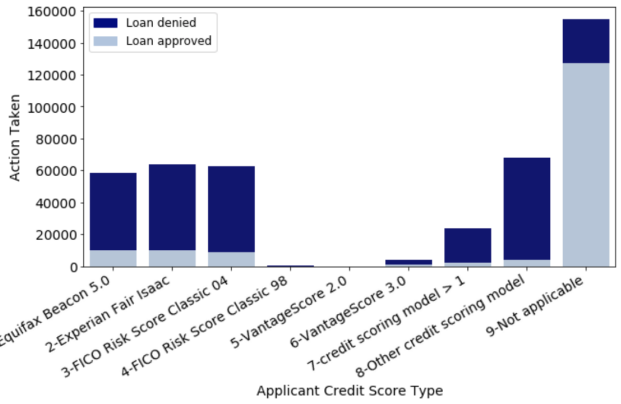

Figure 1.5: Credit-score-type Variable Analysis

2 Feature Engineering

2.1 Categorical Variable Encoding

In order to prepare for feature selection and modeling process, we need to complete feature engineering for different kinds of variables. We apply one-hot encoding techniques

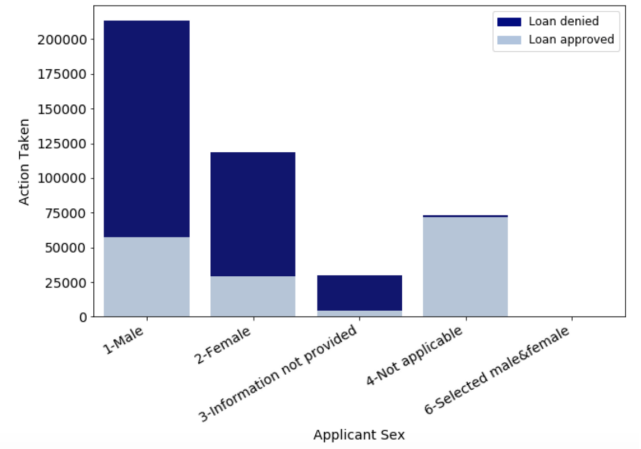

Figure 1.6: Sex Variable Analysis

for all the nominal variables where

$$X = \{stage1, stage2, stage3\}$$

and

$$\varphi(x) = [\mathbb{I}(x = stage1)\mathbb{I}(x = stage2)\mathbb{I}(x = stage3)]\,.$$

As for ordinal variables, we employ Real Encoding methods, where

$$X = \{level1, level2, level3\}$$

and

$$\varphi(x) = [\mathbb{I}(x \geq level1)\mathbb{I}(x \geq level2)\mathbb{I}(x \geq level3)]\,.$$

This approach will generate n-1 dummy variables to represent the information of n-level ordinal variable.

2.2 Numerical Variable Normalization

For numerical variables, we apply the normalization approach to scale all the variables from different ranges to the similar ones with mean = 0 and standard deviation = 1. In this case, we could reduce the negative impact of the multi-collinearity problem, enhance the interpretability of our models, and make it easy to compare different feature importance to do feature selection as well as model analysis.

3 Model Selection

3.1 Feature Selection

3.1.1 Pearson Correlation Approach

In this approach, we consider numerical variables and category variables separately.
For categorical variables, after we apply the one-hot encoding and real encoding methods to nominal as well as ordinal variables, we compute the correlation coefficient between all the dummy variables and our target variable "action-taken" following the equation below:

$$r = \frac{\sum_{i=1}^{n}(Xi - X)(Yi - Y)}{\sqrt{\sum_{i=1}^{n}(Xi - X)^2}\sqrt{\sum_{i=1}^{n}(Yi - Y)^2}}$$

After this, we select in total 20 categorical dummy variables in terms of the absolute value of their correlation with label y whose correlation coefficients are larger than 0.2.

As for numerical variables, similarly, we compute their correlation with target variable and select the top 5 numerical variables whose correlation coefficients are larger than 0.1.

3.1.2 Logistic regression with L1 regularization Approach

Employing logistic regression method with L1 regularization(Lasso) method, we plan to make use of the sparsity property of its coefficient due to its L1 regularization, to complete the feature selection. The underlying principles of Logistic regression with L1 regularization term are as follows:

$$w^{optimal} = \min_{w \in R} \sum_{i=1}^{n} log(1 + e^{-y_i \mathbf{X_i}^\top w}) + C\,\|w\|_1$$

Figure 3.1 is a basic illustration about the sparsity property of Lasso method, where the left plot is ridge and the right plot is lasso, and the solid areas are the constraint regions, while the red ellipses are the contours of sum of square in regression case.

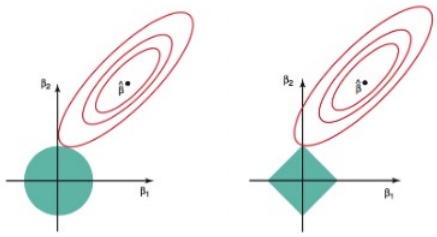

Figure 3.1: Lasso Feature Selection Principle

We include all the numerical variables as well as dummies from categorical variable encoding in to our logistic regression with L1 regularization term model, and select corresponding numerical variables and dummy variables whose coefficient in this model is significantly nonzero. As a result, we select 62 predictors including 1 numerical variable "income" and 61 dummy variables.

3.2 Model Selection

3.2.1 Logistic Regression Approach

After we select important features based on two different methods, we can move on to our model training and selection section. Firstly, we apply logistic regression with L1 regularization term model for our dataset under two feature selection methods. The specific principle of this model is that firstly, we follow the same formula shown in Lasso feature selection method section to get optimal w. Secondly, we could compute probability of the predicted yi belonging to class 1(approval) given Xi, and set the threshold as 0.5:

$$P(\hat{y} = 1) = \frac{e^{\mathbf{X_i}^\top w}}{1 + e^{\mathbf{X_i}^\top w}}$$

$$\hat{y} = \begin{cases} 1 & ,if P(\hat{y} = 1) \geq 0.5 \\ 0 & ,if P(\hat{y} = 1) < 0.5 \end{cases}$$

In addition, as for this logistic regression model, we will tune the penalty coefficient C through cross-validation approach to obtain the optimal logistic regression model.

As for our first feature selection method, Pearson correlation approach, we employ the logistic regression model and apply the 10-fold cross-validation to obtain its model accuracy. The tuning process of penalty coefficient C and its optimal model ROC curve are as follows in Figure 3.2.

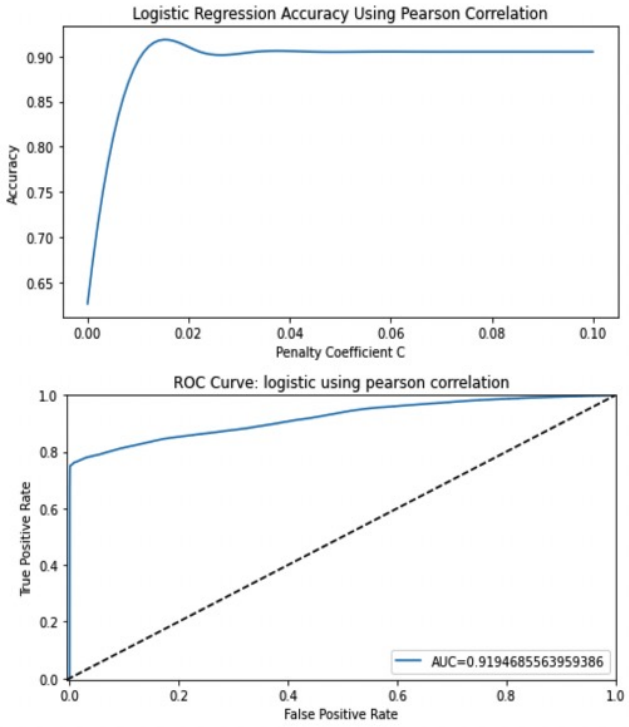

Figure 3.2: Tuning Process and ROC Curve—-Logistic Regression with Pearson Correlation Feature Selection Method

As a result, the average model accuracy from cross-validation reaches 0.9051 for the Logistic Regression Approach after Pearson's correlation feature selection.
As for the second feature selection method, lasso feature selection, since the modeling methods is the same as feature selection, we just make use of the model in lasso feature selection process, and apply the 10-fold cross-validation to find its optimal model parameter. The tuning process and its corresponding final ROC curve for optimal model is shown in Figure 3.3. The results turn out that the average model accuracy is 0.9048, which is relatively lower than first feature selection method.

3.2.2 Random Forest Approach

Secondly, we use random forest approach under two feature selection methods separately. The principle about random forest is that it combines groups of decision trees where each decision tree include randomly assigned groups of features to train. As for each tree, it will be split based on reducing the largest amount of entropy at each step to converge. After this, when we make prediction, we will take the majority vote from all decision trees to predict classes of each input sample. Figure 3.4 shows the basic principles how Random Forest model works.[1]

For this modeling approach in our case, we include 100 trees in the forest, and regard the maximum number of depth for each tree as our tuning parameter. Specifically,

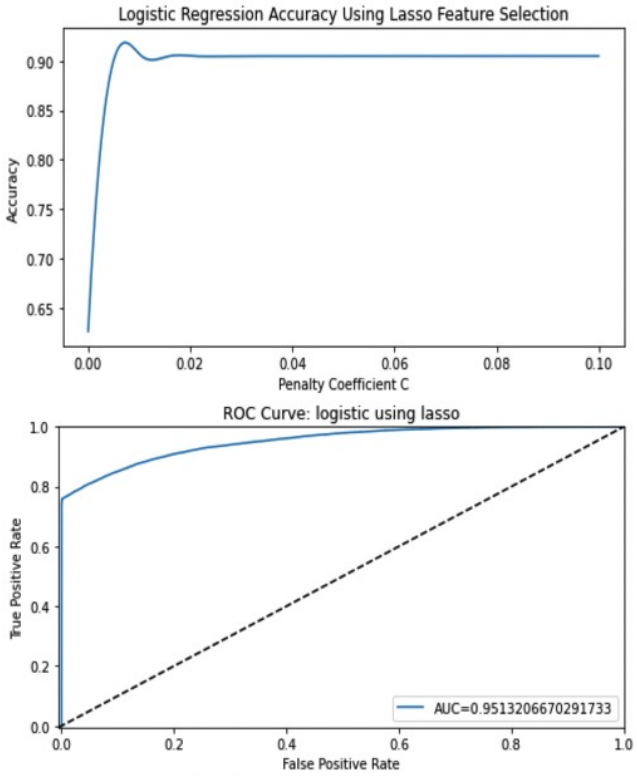

Figure 3.3: Tuning Process and ROC Curve—-Logistic Regression with Lasso Feature Selection Method

we will tune the maximum number of depth in each tree through grid search and cross-validation to obtain its optimal random forest model.
Similar as logistic regression training process, we apply the 10-fold cross-validation to obtain its model accuracy. The average model accuracy is 0.9091 for the correlation feature selection approach, which is a bit higher than logistic regression model, and the tuning process to find optimal parameter and corresponding model is shown in Figure 3.5.

As for the second feature selection approach, the results turn out that the average model accuracy is 0.9102, which is the highest one among all the combination of models and feature selection methods. the tuning process to find optimal parameter and corresponding model is as follows in Figure 3.6.

3.2.3 Model Comparison

After we have employed logistic regression models as well as random forest models together with two different feature selection methods, Pearson correlation and Lasso, we could make a comparison for the four optimal models under each combination of feature selection and modeling approaches. Table 3.1 summarizes the performance of two models under different feature selection approaches.
As we can see from the Table 3.1, the current optimal model among four is the random forest model with logistic regression with L1 feature selection methods, which gives an average accuracy of 0.9102. However, the average accuracy of all four models are close, meaning that tour feature selection and modeling approaches are all quite fitted with our data.

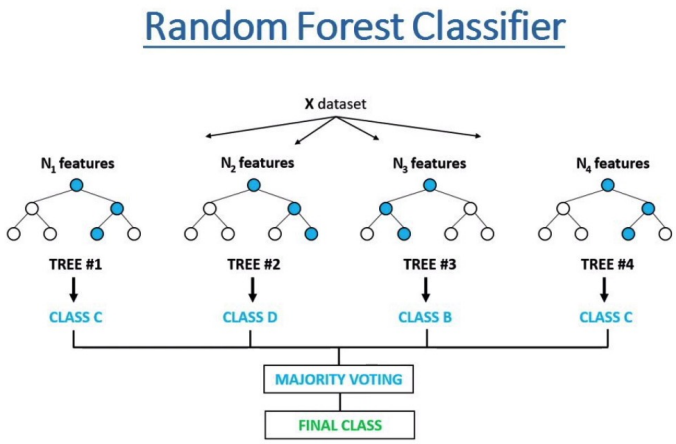

Figure 3.4: Random Forest Model Principle

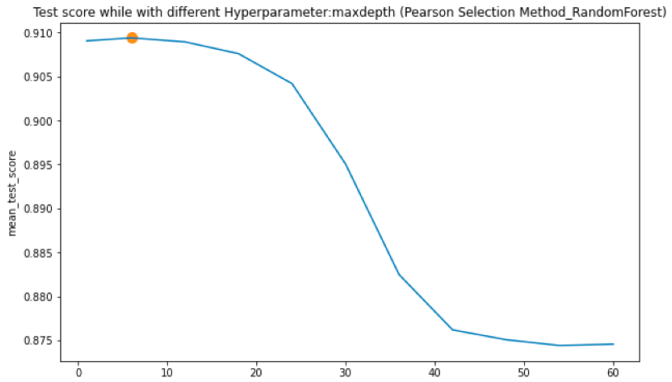

Figure 3.5: Tuning Process—-Random Forest with Pearson Correlation Feature Selection Method

As for explanation for this results, firstly, as for feature selection part, since the Pearson correlation method only consider the paired relationship between certain variable and label but Lasso method consider relationship between all the variables and label, which could give us a better subset of features and a better accuracy for model prediction. Secondly, as for the model selection, since random forest is much more complex than logistic regression, and after our feature selection, we have reduced the overfitting issues to a large extent, the optimal model from random forest should in general perform better then logistic regression.

3.2.4 Model Improvement: Interaction terms + PCA

In this section, we will further consider some model improvements methods to try to increase model prediction accuracy.

We firstly consider to include two-dimension interaction terms of our selected features after feature selection step into our model ing process, which may help increase the model complexity and interpretability.

However, since there may exist overfitting issues after we include all the two-dimension interaction terms, we further apply the PCA method to reduce the dimension and relieve the overfitting problem. In specific, we want the explained variance after PCA to be 95%, thus, a cut-off threshold of 95% is set to choose the number of principle components in

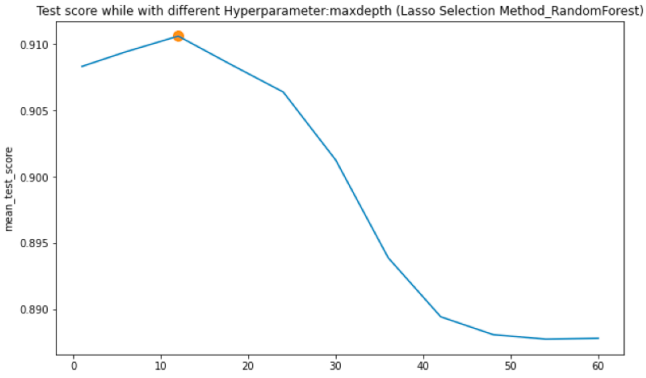

Figure 3.6: Tuning Process—-Random Forest with Lasso Feature Selection Method

Feature Selection Methods	Average Model Accuracy	
	Random Forest	Logistic Regression
Pearson Correlation	0.9091	0.9051
Logistic Regression with L1 (Lasso)	0.9102	0.9048

Table 3.1: Model Accuracy Comparison

PCA approach, and Figure 3.7 shows the process of number of components selection in the logistic regression model with Pearson correlation method.

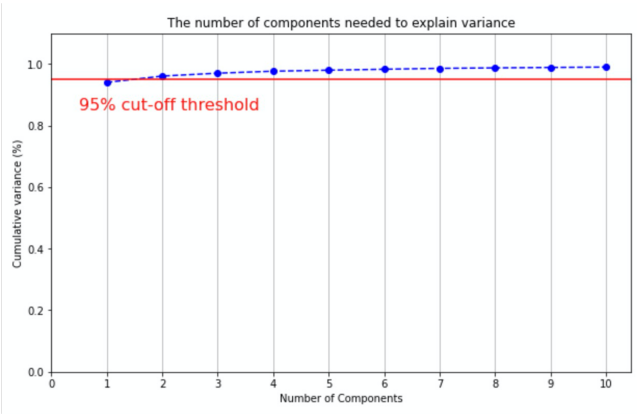

Figure 3.7: PCA Number of Components Selection

Similar as before, in this section, we will consider the logistic regression and random forest modeling approaches together with two different feature selection methods, and we will also apply 10-fold cross-validation to obtain the optimal model parameters under each approach and their corresponding optimal prediction accuracy.

The Table 3.2 shows the optimal model accuracy comparison of in total the four models after including two-dimensional interaction terms and PCA methods. However, we find that the accuracy under each combination of two feature selection methods and two models are all smaller than models without interaction terms and PCA methods.

The potential rationale for this phenomenon is that we have considered all the two-dimension interaction terms into our

model, which increase the co-linearity between all the input variables. In this case, if we then apply PCA approach to reduce the dimension to get certain number of principle components, we may lose some information about variable variance due to co-linearity in our whole predictors matrix. Also, the variables of high correlations will be loaded out on the same Principal Component (Eigenvector). Thus, PCA approach may not be a perfect solution to the curse of dimensionality of this new dataset. Thus, under this circumstance, we decide to return to our previous models without considering interaction terms and PCA methods to obtain our final optimal model.

Feature Selection Methods	Average Model Accuracy	
	Random Forest	Logistic Regression
Pearson Correlation	0.9036	0.9041
Logistic Regression with L1 (Lasso)	0.8976	0.8848

Table 3.2: Model Accuracy with Interaction Terms + PCA

4 Further Discussion

4.1 Feature Importance Analysis

After we have selected the overall optimal model, which is the Random Forest with Lasso Feature Selection Method without interaction terms, we further conduct analysis to explore its feature importance issues.

In Figure 4.1, when we show the top 15 important features from the optimal model, debt-to-income ratio, hoepa-status, and purchaser-type is the three most crucial features to make approval or denial decisions in model, where hoepa-status refers to whether the covered loan is a high-cost mortgage. Moreover, the initially-payable-to-institution, the age, race, and sex of applicant and co-applicant are also significant features accounting for the approval or denial decision prediction.

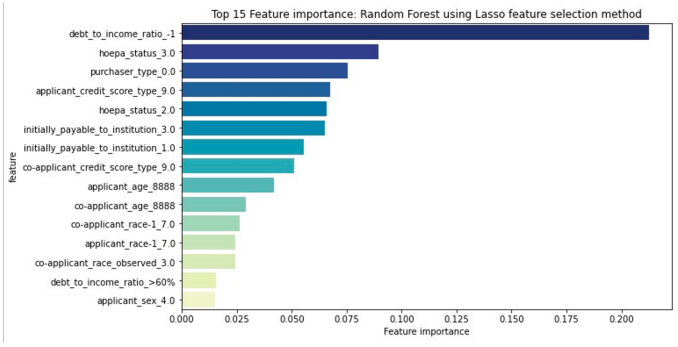

Figure 4.1: Feature Importance of Optimal Model

4.2 Fairness Analysis

The machine learning algorithms are affecting financial decisions where discrimination and bias are possible to occur. Especially in our case, the approval of loans is a significant social and financial activity in a society, also, for individuals. Thus, in terms of fairness, equality should be addressed to avoid oppression and bias on specific groups of people. Specifically, we do not anticipate any racial and gender information will have any impact on our decision.

However, according to results in Table 4.1, we find that race and sex are both important features in model prediction, which may contradict with the view from fairness that groups from these two variables should be protected from discrimination. Therefore, we follow the similar procedures as model selection section to obtain the optimal models after eliminating the sex or race features separately, and then compare these two optimal model accuracy with the models including sex or race features. Our model comparison summary are as follows in Table 4.1.

Protected Attribute(X)	Accuracy from Optimal Model	
	Model with X	Model without X
Sex	0.9102	0.9102
Race	0.9102	0.9099
Sex and Race	0.9102	0.9098

Table 4.1: Fairness Analysis

From Table 4.1, we could conclude that after eliminating protected attributes "sex" or "race" or "sex and race" from our original dataset, we still obtain an optimal model with approximately the same accuracy as before. Thus, we can obtain a final optimal model achieving good fairness and accuracy simultaneously by eliminating sex and race features from predictors. This model could be our final prediction model for financial institutions to decide whether to offer credit to a client.

4.3 Limitations and Improvements

4.3.1 Completeness of Data

Among all 98 predictors, there are more than 30 features that have NA proportion larger than 50%, where we have to conduct separate strategies to fill in these values, drop variables or delete corresponding rows based on their NA proportion number. Therefore, if we could have a dataset with much smaller proportion of NA values, we may give out a model with better interpretability and prediction performance.

On the other hand, as previously stated, the important features contain some race, ethnicity and gender information, however, they do not cause any problems on fairness. The reason is that the dummy variables after encoding that are shown in Figure 4.1, are "not applicable" values. It implies that it is the incompleteness that causes the significance rather than their true gender or race class information. In fact, it can be interpreted that having complete information or not is a significant feature. Thus, if we have a more complete original dataset, we may come out with a different interpretation on these social-related features from models.

4.3.2 Feature Selection Methods

In feature selection section, we employed the Pearson correlation approach and Logistic regression with L1 regularization Approach (Lasso) approach to help select important features for modeling process. However, these two methods

may not be the optimal ones in terms of different modeling approaches.

Therefore, we could apply better feature selection like backward step-wise subset selection algorithms with relative importance as the measure[2]. For this method, we will start with all features, and at each step, we will train the corresponding models(random forest or logistic regression) and delete the least relevant feature with the smallest relative importance(R_a), which can be computed in the following way:

$$V_a = \sum_{j=1}^{L} (\hat{y_{a_j}} - \bar{\hat{y_{a_j}}})/(L-1)$$

$$R_a = V_a \sum_{i=1}^{I} V_i * 100\%$$

The previous process will be repeated until the stopping criterion is met. Overall, through this approach, we could make use of specific training model to select its best subset of features, meaning that it is a much more model-oriented method, which may give a better model prediction accuracy in the end.

4.3.3 Dimension Reduction Methods

In our model improvement section, we try to include interaction terms and apply PCA dimension reduction method to alleviate the overfitting issues. However, since our whole input features including interaction terms may have colinearity issues, PCA may results in a small number of principle components and lead to a loss of some important feature information for modelling process.

As for this problem, we can try other dimension reduction methods like "Locally linear embedding(LLM)" approach or "Laplacian Eigenmaps" approach after including all two-dimensional interaction terms in our model, to help reduce overfitting issues. In this case, we may obtain a better model prediction accuracy considering interaction terms than our current optimal model.

5 Conclusion

Predictions on whether to offer credit to a client is of great significant for financial institutions to greatly reduce the default probability. In our project, we firstly apply Pearson correlation and Lasso feature selection methods to select important features to reduce overfitting issues and increase model interpretability. After this, we employ logistic regression and random forest approaches for modelling process. Among the four optimal models under each feature and model selection combination, the random forest based on Lasso selection approach gives the best prediction accuracy, which is 91.02%.

We are confident that our model can be applied to other new datasets since random forest is a flexible model by adjusting its maximum tree depth or number of features considered in each split to fit new data. Moreover, we also employ feature selection approach to reduce its overfitting problems, in order to improve its prediction accuracy on new data. In addition, as for our final model, we may eliminate the sex and race features from our optimal model set to reach model fairness and accuracy simultaneously.

Lastly, for our final model to be more informative and valuable, further improvement needs to be performed. Firstly, lack of data completeness is limitation. We expects the model to predict better if the NA data proportion in original dataset could decrease by a large amount. Secondly, we will also consider more advanced feature selection approaches and dimension reduction methods to alleviate overfitting problems and improve model intepretability as well as prediction accuracy. Overall, we hope this model could be a great tool to benefit the financial institutions and credit market.

6 References

[1] Image obtained from Medium, Chauhan, A. (2021, February 23). Random Forest classifier and its hyperparameters. Medium. Retrieved December 6, 2021, from https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6.

[2] Cortez, P., António Cerdeira, Almeida, F., Matos, T., and José Reis. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553