ORIE 5741 Data Analysis Project

Credit Approval Prediction

Ruifan Chen
Xinyue Liu
Yijun Zhu
Dec 5, 2021

Content

1. Introduction
2. Data Preprocessing
3. Data visualization
4. Feature Engineering
5. Model Selection
6. Further Analysis

Introduction: Data Description & Project Goal

HMDA home mortgage disclosure action Data

+

Median Housing price data

86 categorical variables

11 numerical variables

1825608 samples

- financial and demographic background
- demographic information of the county or tract
- Information about the loan like loan term, loan amount

Goal of Project:

1. Exploratory data analysis incorporating our understanding of feature importance of approval decision

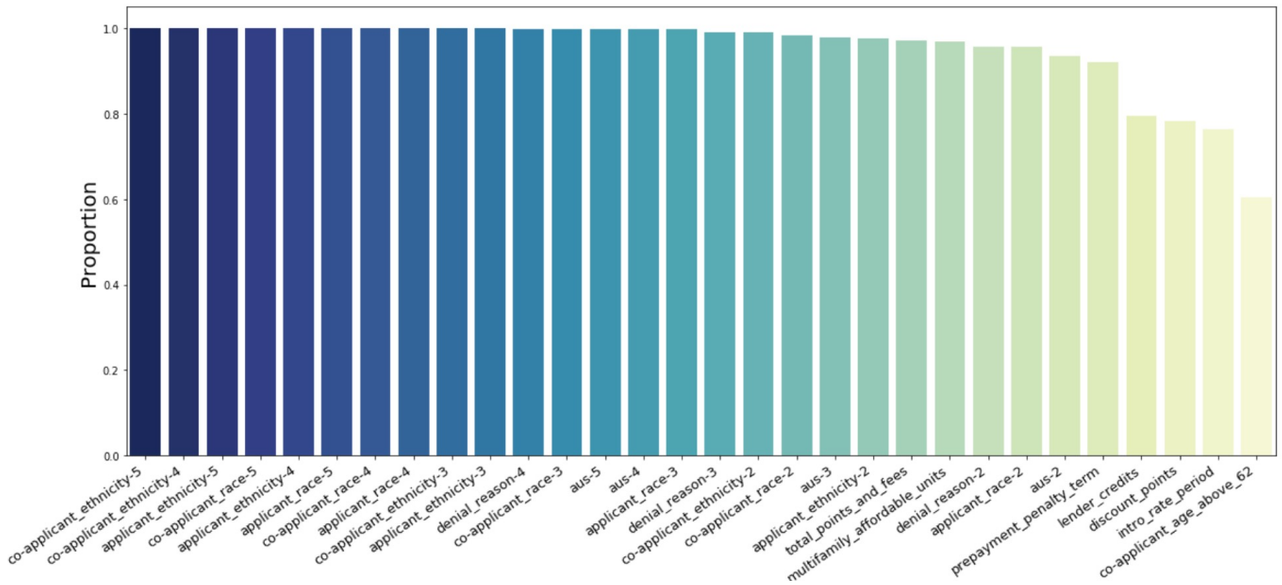
2. Construct a classification model to make better decisions on whether to offer credit to a client

Content

1. Introduction
2. Data Preprocessing
3. Data visualization
4. Feature Engineering
5. Model Selection
6. Further Analysis

Data Preprocessing : NA values, Outlier Detection

NA values



- More than 30 variables have a proportion of over 50 percent of null values.

1. Variables with over 50 percent of NA values:

   Drop variables

2. For those variables with 5 to 50 percent of NA values:

   Fill in with estimate by random forest models.

3. Variables with less than 5 percent of NA values:

   Delete corresponding rows.

Outlier Detection

Interquartile Range Approach

$$(\boldsymbol{Lower\ Quantile\ -1.5\ *Interquartile\ Range},\ \boldsymbol{Upper\ Quantile+1.5\ *Interquartile\ Range})$$

Note:

- Lower Quantile(Q1) is 25% Quantile,
- Upper Quantile(Q3) is 75% Quantile,
- Interquartile Range = Q3 - Q1

Data Preprocessing : Label Simplification

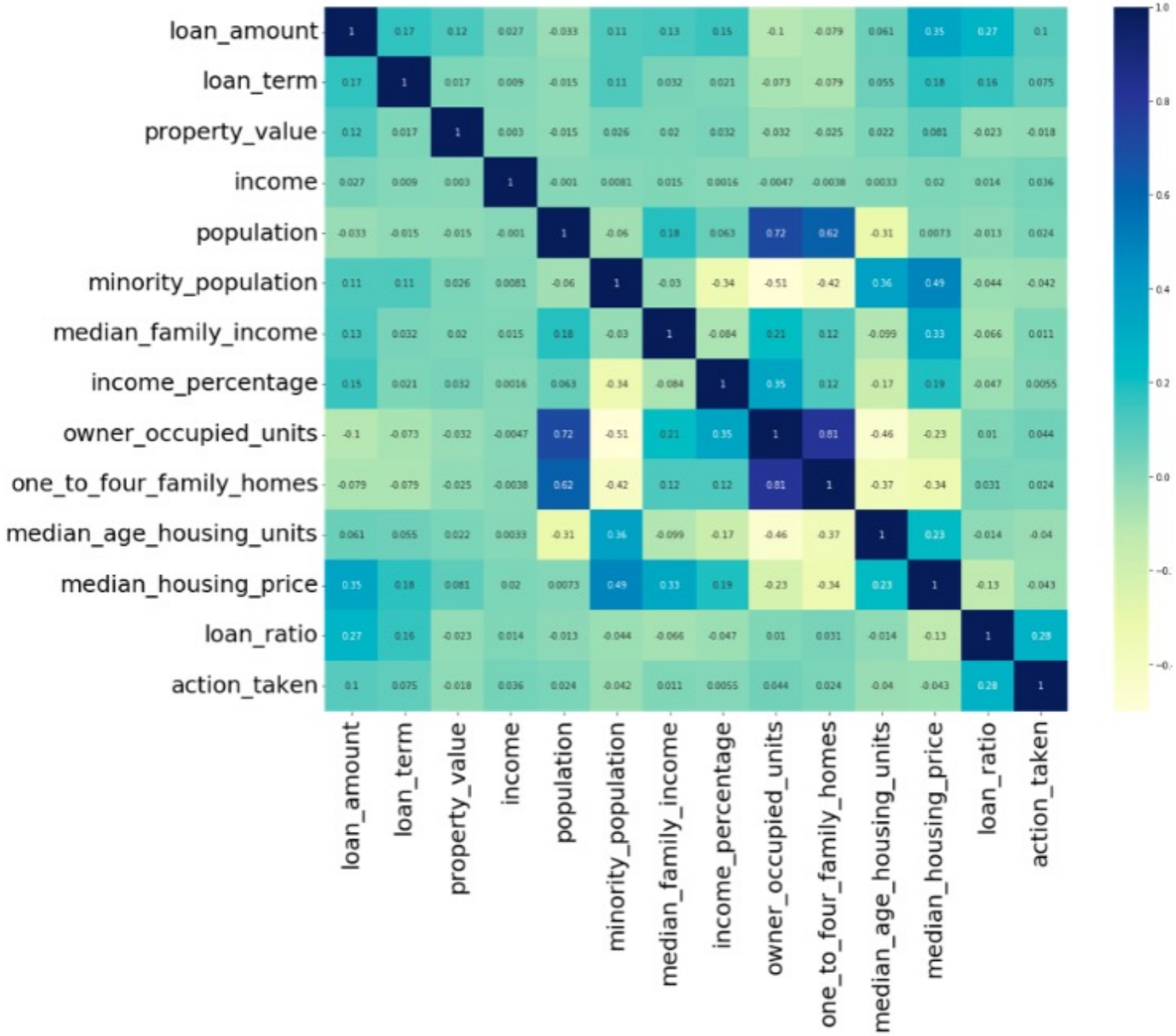
Original Target Label: “action-taken”

- 1 - Loan originated
- 2 - Application approved but not accepted
- 3 - Application denied
- 4 - Application withdrawn by applicant
- 5 - File closed for incompleteness
- 6 - Purchased loan
- 7 - Preapproval request denied
- 8 - Preapproval request approved but not accepted

New Target Label: “action-taken”

- 1 – Approval: “Application approved but not accepted” + “Purchased loan”
- 0 – Denial: “Application denied” + “Preapproval request denied”

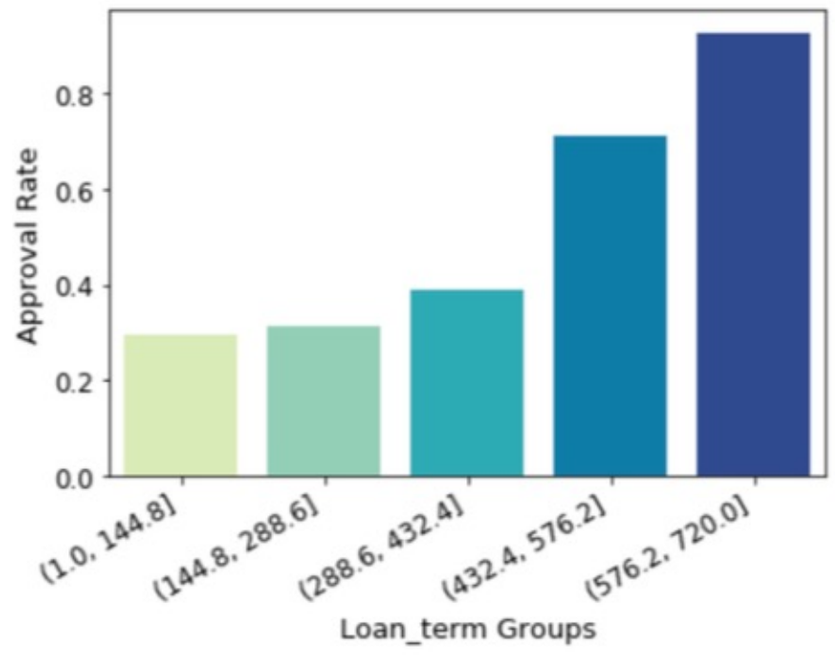
Data Visualization: Correlation Analysis

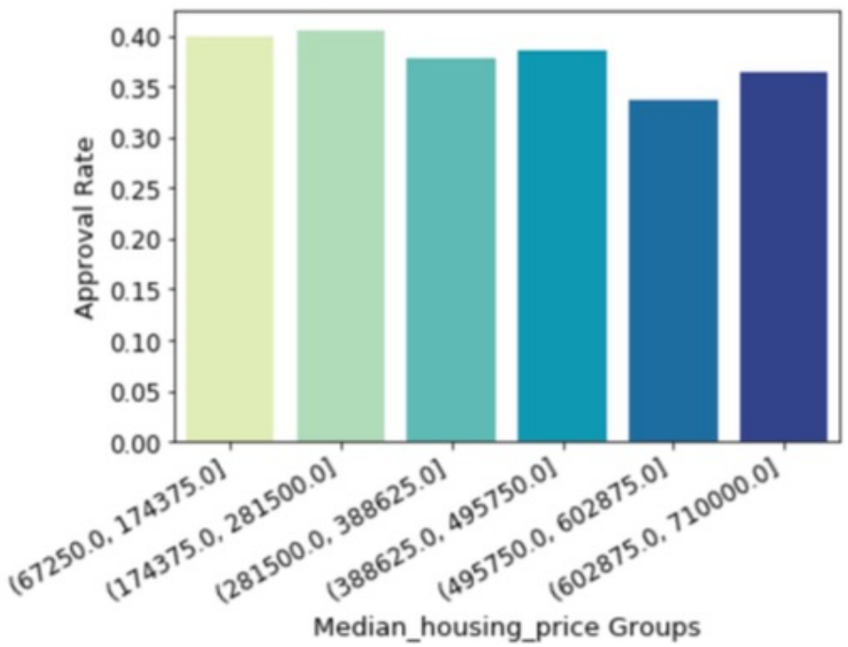


- For numerical ones, we conduct a correlation heatmap to explore the relationship between each variable and the label.

- Label “Action-taken” is relatively highly correlated with features “loan-term” and “median-housing-price”.

Data Visualization: Important Numerical Variable Visualization

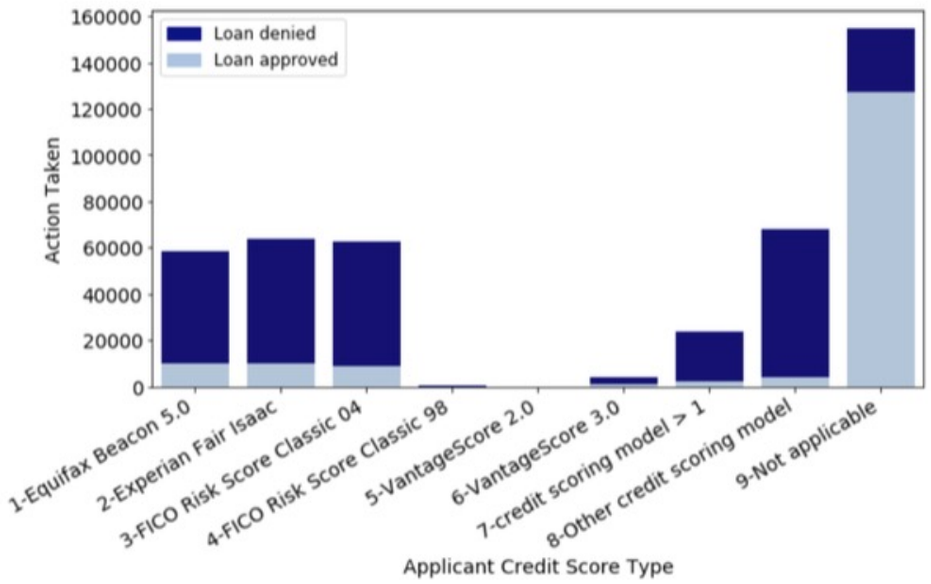


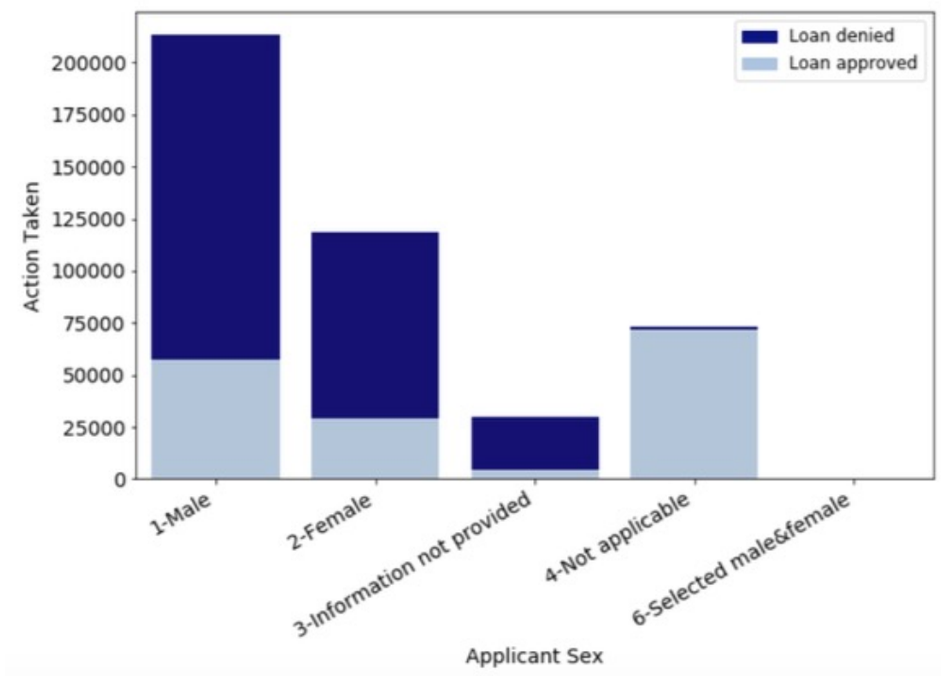


- For numerical variables, we consider the two most correlated numerical variables: “loan-term” and “median-housing-price”, and plot group bar charts to see their influences on the credit approval rate.

- 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”.

Data Visualization: Important Categorical Variable Visualization





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

- We 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.

Content

1. Introduction
2. Data Prepossessing
3. Data visualization
4. Feature Engineering
5. Model Selection
6. Further Analysis

Feature Engineering: Categorical Variable Encoding, Numerical Variable Normalization

1 Categorical Variable Encoding

Nominal Variable: One-hot Encoding

- E.g. X = {level1, level2, level3}
- $\varphi(x) = [\mathbb{I}(x = level1), \mathbb{I}(x = level2), \mathbb{I}(x = level3)]$
- One-hot encoding: only one element is non-zero

Ordinal Variable: Real Encoding

- E.g. X = {Stage1, Stage2, Stage3, Stage4}
- $\varphi(x) = [\mathbb{I}(x \geq level1), \mathbb{I}(x \geq level2), \mathbb{I}(x \geq level3), \mathbb{I}(x \geq level4)]$
- Real Encoding: more than one element could be non-zero

2 Numerical Variable Normalization

$$Xs = \frac{X - mean(X)}{sd(X)}$$

$$\text{ys} = \frac{y - mean(y)}{sd(y)}$$

- Reduce the multi-collinearity effect detected by computer
- Convenient for feature importance comparison and interpretability

Content

1. Introduction
2. Data Prepossessing
3. Data visualization
4. Feature Engineering
5. Model Selection
6. Further Analysis

Model Selection: Feature Selection

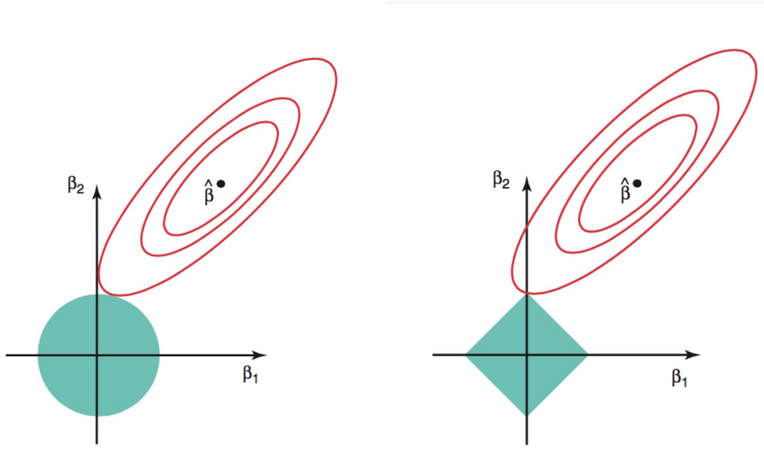
1 Pearson Correlation Method

$$\rho_{xy} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X,Y)}$$

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

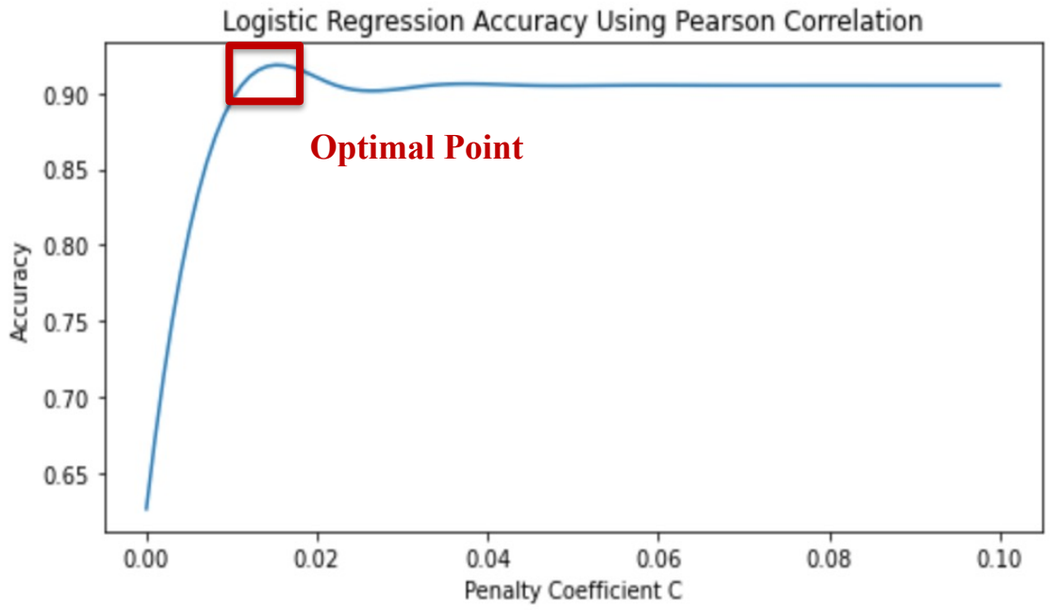
- Compute the correlation coefficient between all the variables and our target variable “action- taken” .
- Select the top 20 categorical and numerical variables separately in terms of their correlations with label y

2 Logistic Regression with L1 Regularization (Lasso Method)

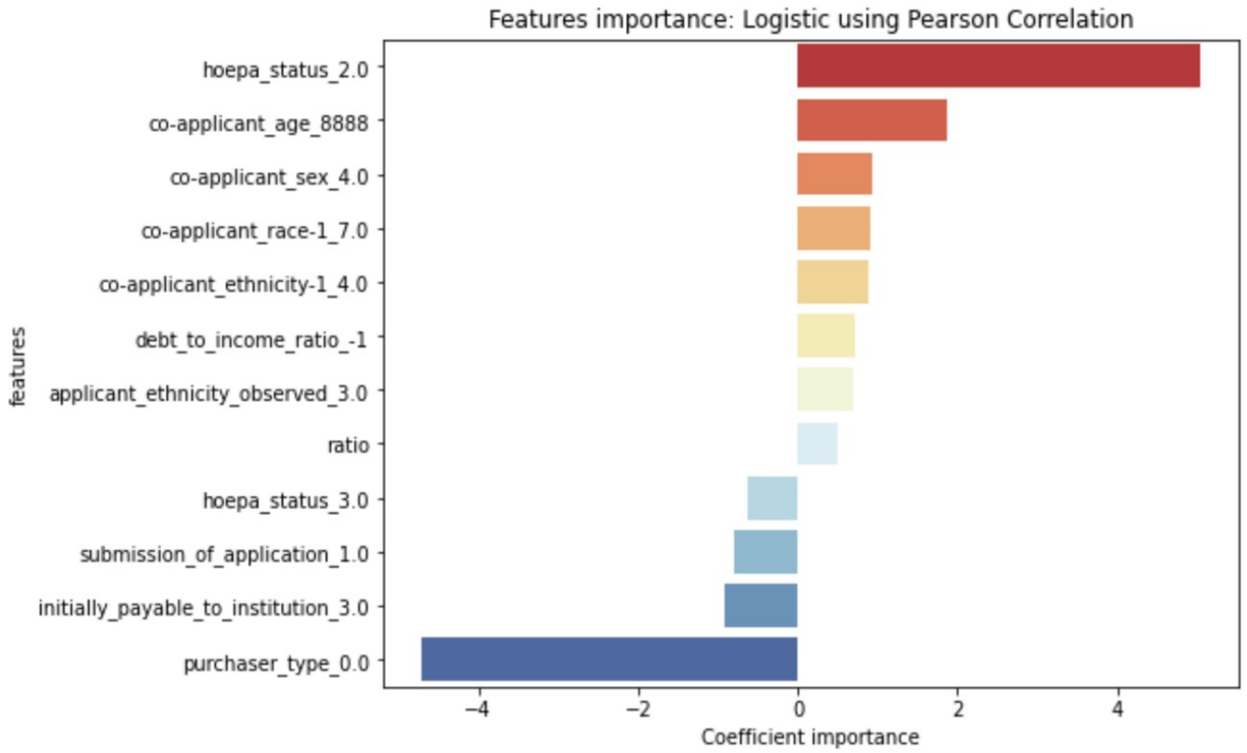


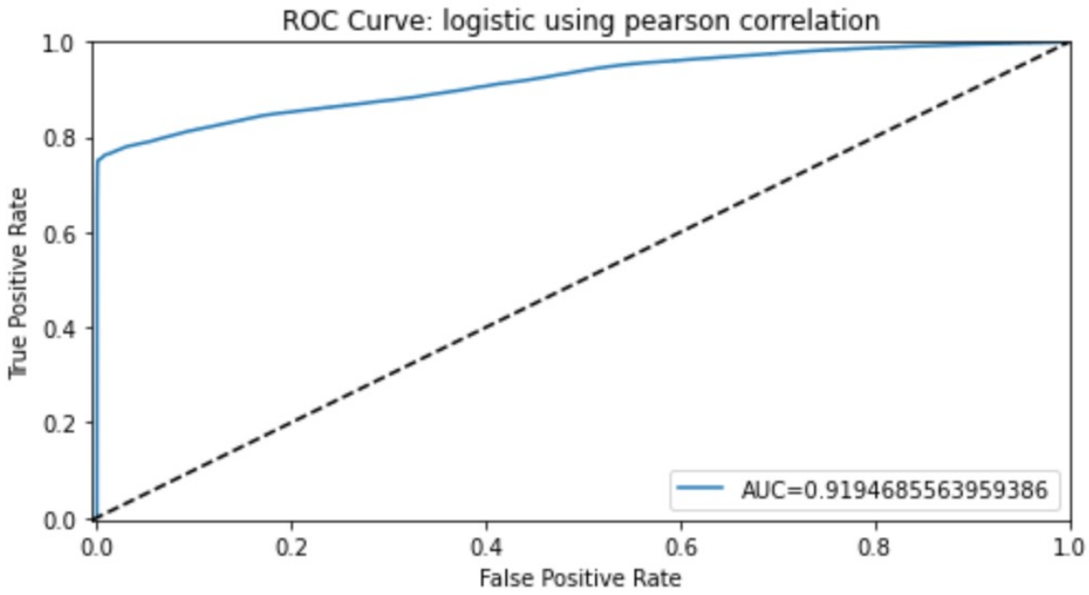
- Consider corresponding numerical variables and dummy variables whose coefficient is larger than 0.01.
- As a result, we select 62 predictors containing 1 numerical variable ”income” and 61 dummy variables.

Logistic Regression Models: Pearson Correlation Feature Selection Approach

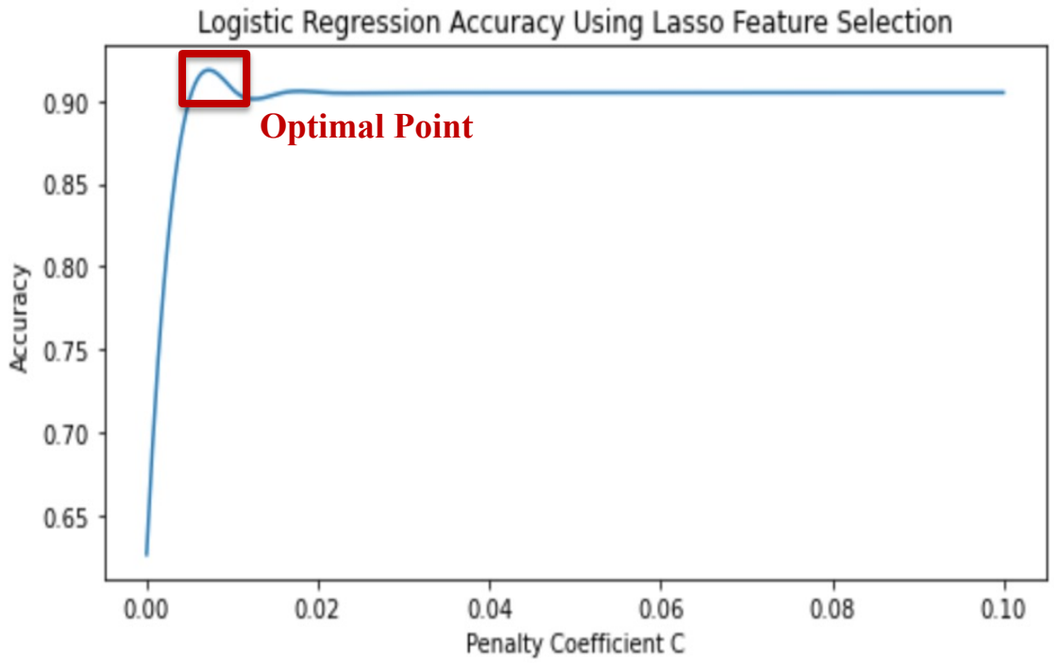


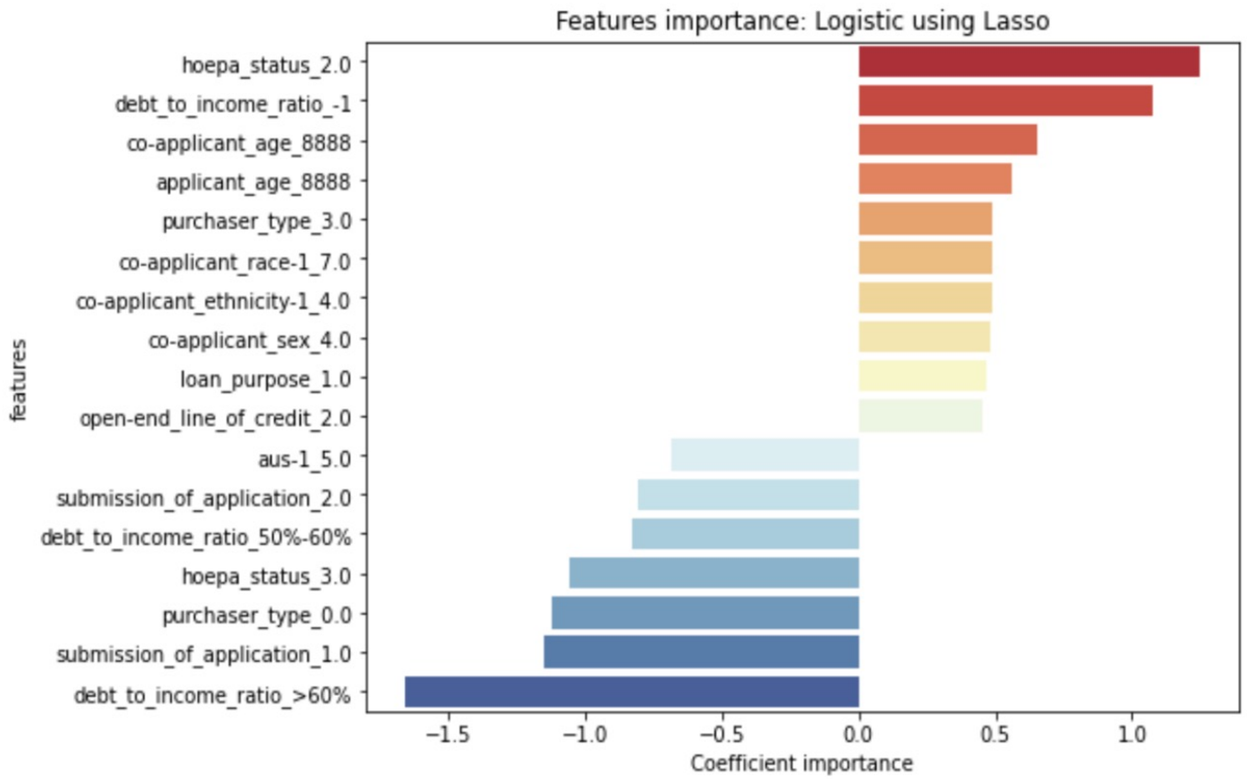
- **Optimal model with accuracy 90.5% and AUC 0.919**
- **Feature importance**
  - Whether the covered loan is a high-cost mortgage
  - Purchaser type: Type of entity purchasing a covered loan



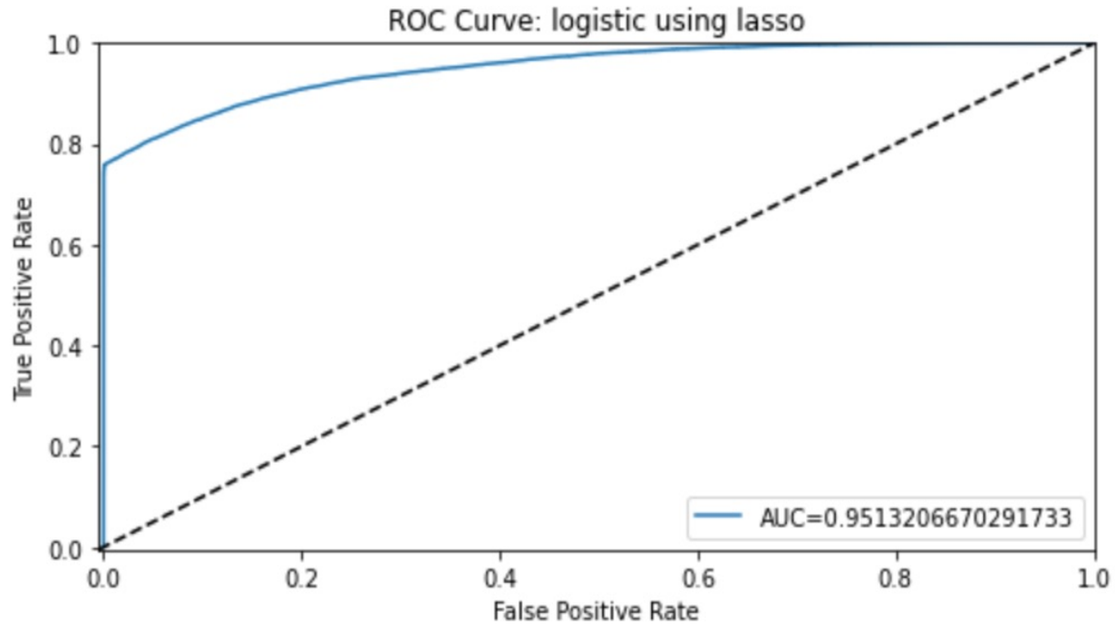


Logistic Regression Models: Lasso Feature Selection Approach



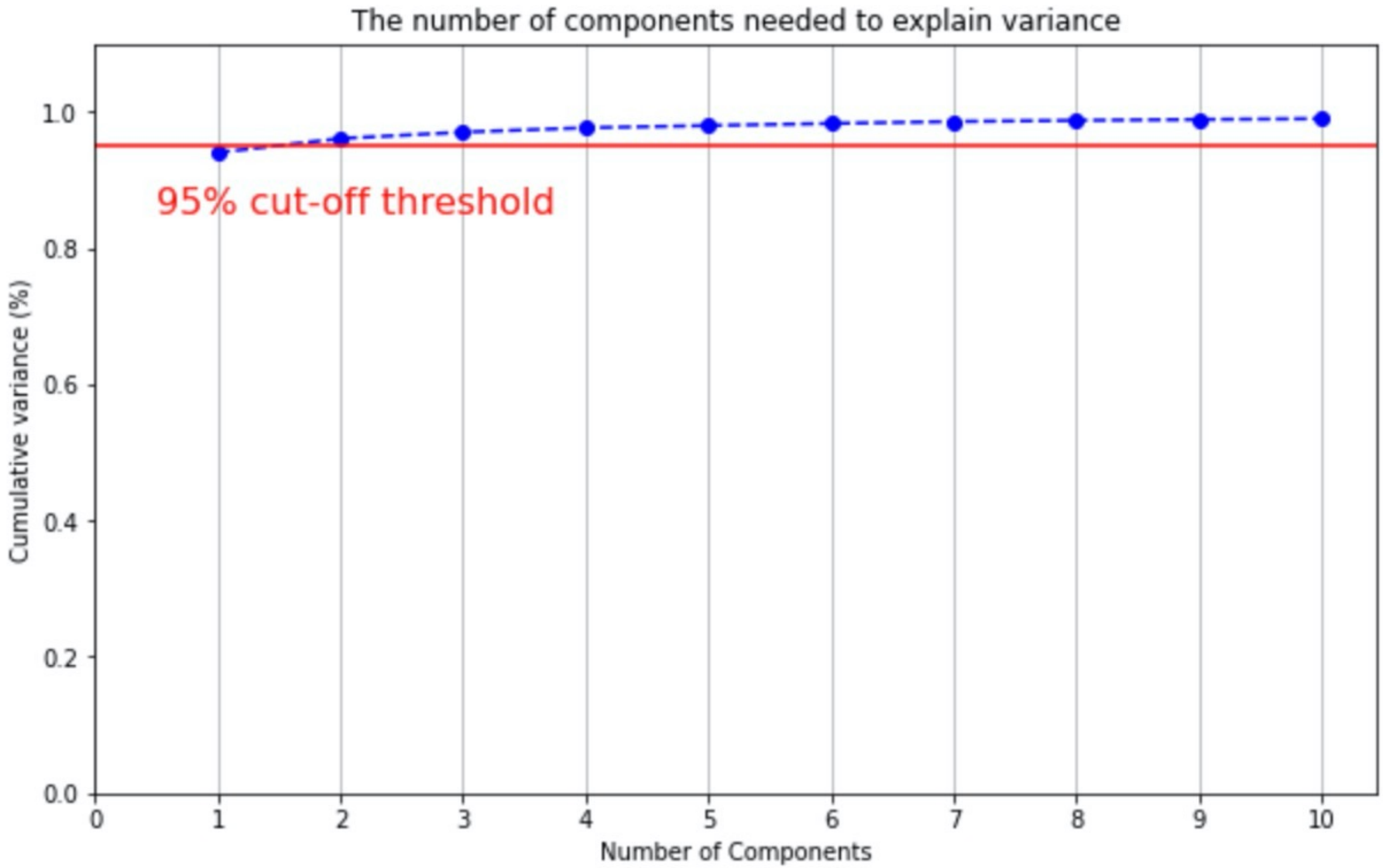


- **Optimal model with accuracy 90.48% and AUC 0.951**
- **Feature importance**
  - Whether the covered loan is a high-cost mortgage
  - Debt-to-income ratio



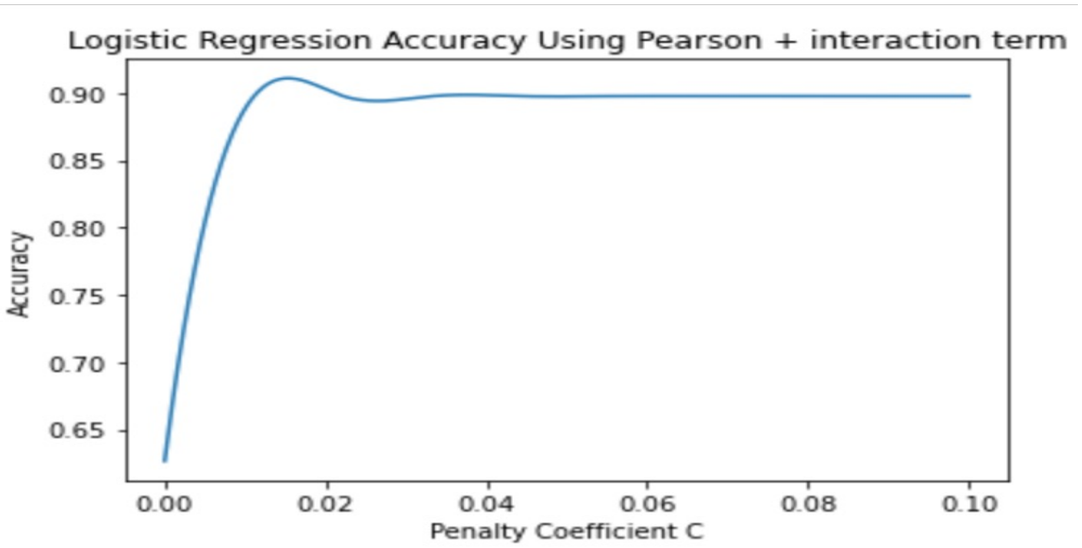
Logistic Regression Models: Interaction Terms + PCA

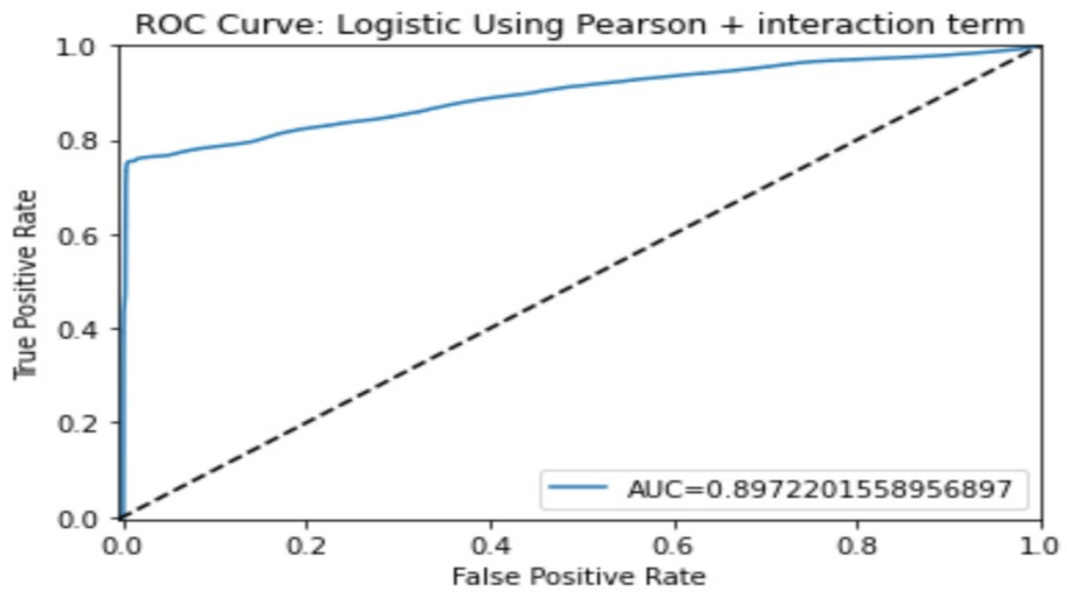
PCA Number of Components Selection



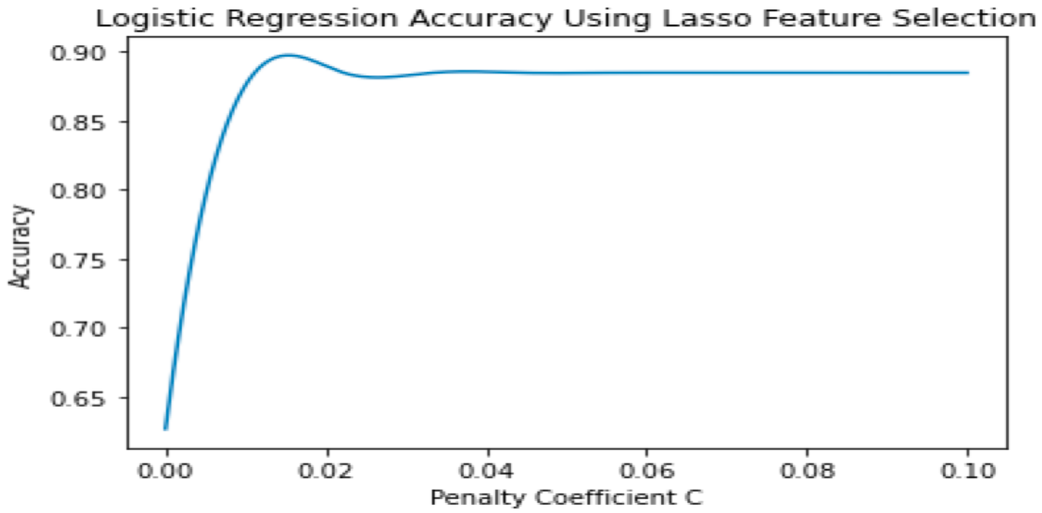
Logistic Regression Models: Interaction Terms + PCA

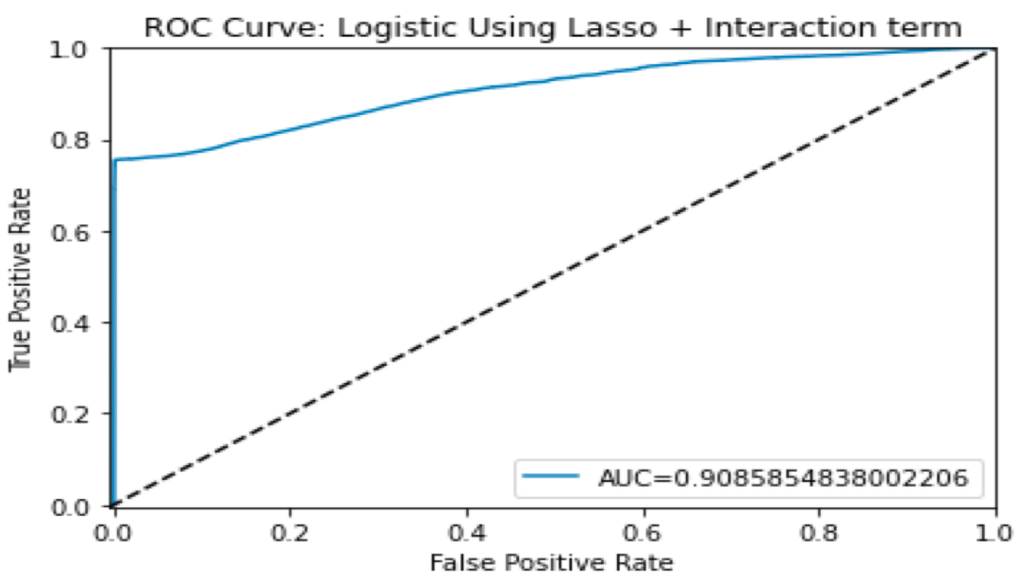
Optimal model using Pearson has accuracy 89.7%





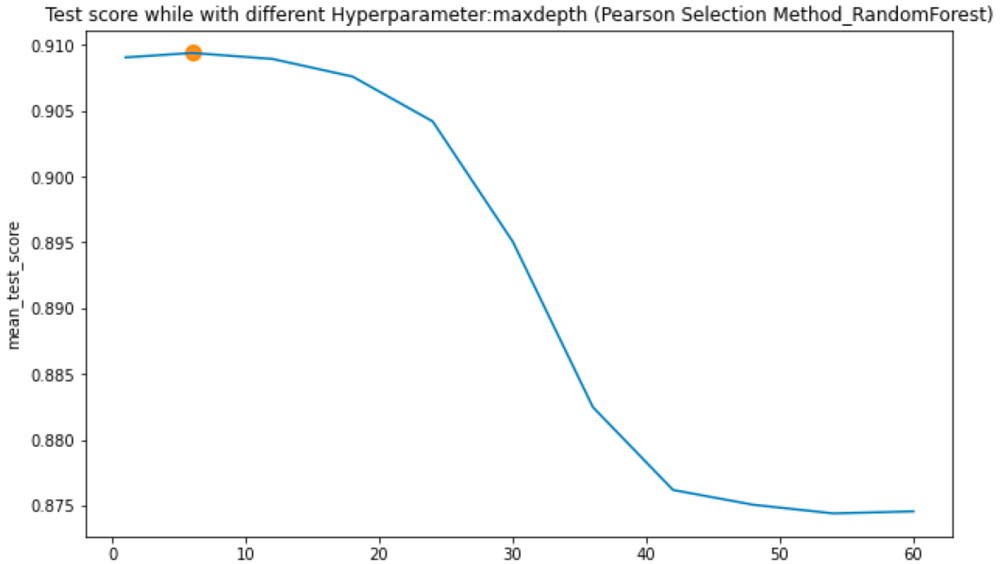
Optimal model using Lasso has accuracy 88.5%

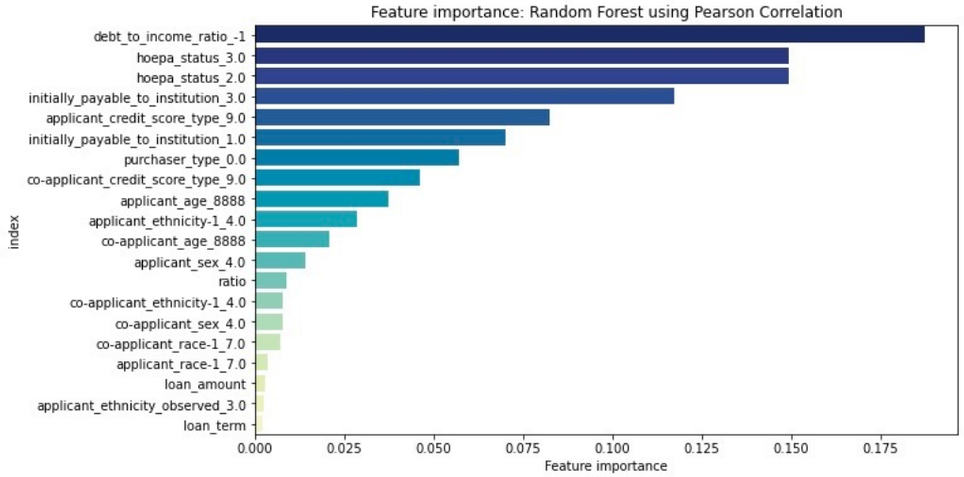




Random Forest Models

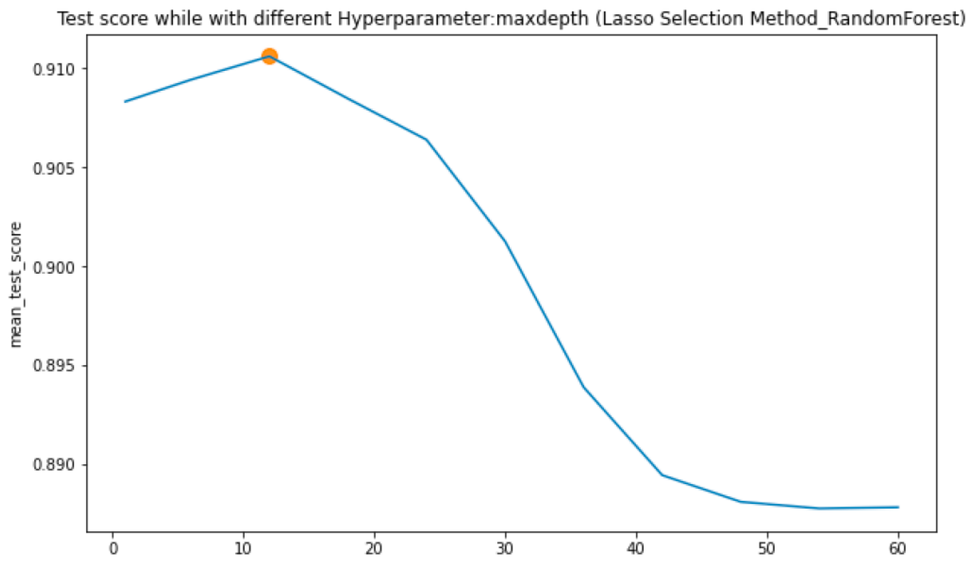
1 Pearson Feature Selection Method

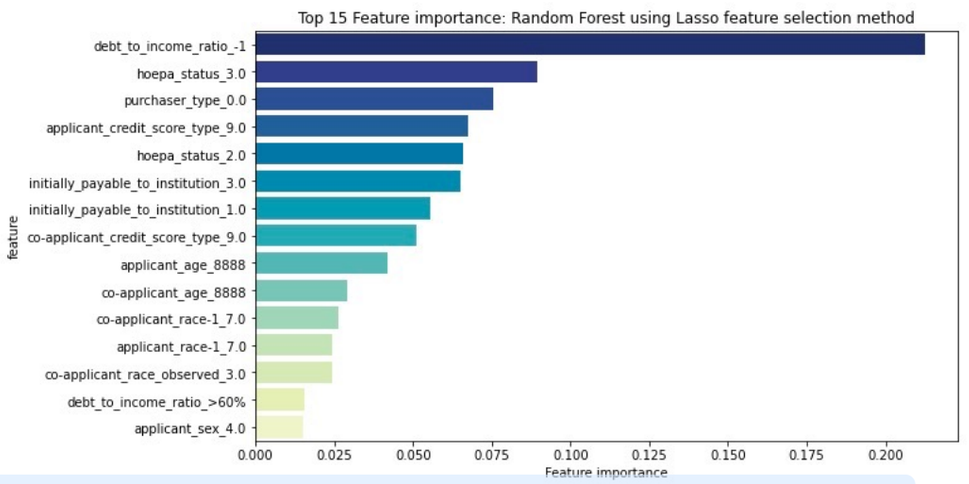




Optimal model using Pearson has accuracy 90.91%

2 Logistic Regression with L1 Feature Selection Method

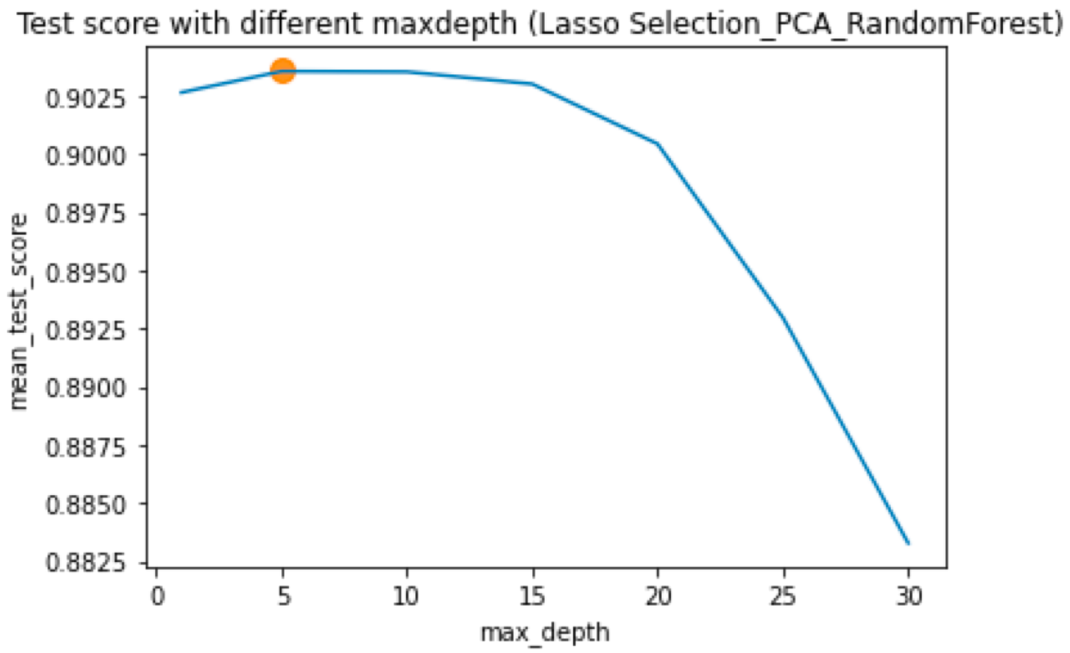




Optimal model using Pearson has accuracy 91.02%

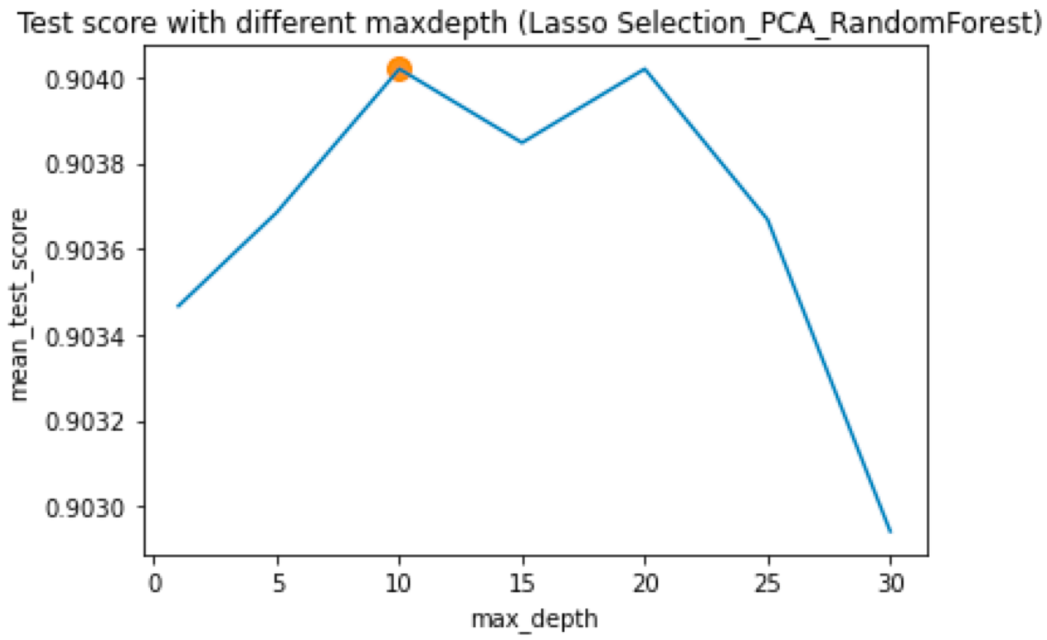
Random Forest Models: Interaction Terms + PCA

Pearson Correlation Selection Method



Optimal model using Pearson has accuracy 90.36%

Logistic regression with L1 Selection Method



Optimal model using Lasso Selection has accuracy 90.41%

Model Selection & Summary

Model	Original Features		Interaction Terms with PCA		Feature Selection
	Pearson	Lasso	Pearson	Lasso	
Random Forest	90.91%	**91.02%** ✓	90.36%	90.41%	
Logistic	90.51%	90.48%	89.76%	88.48%	

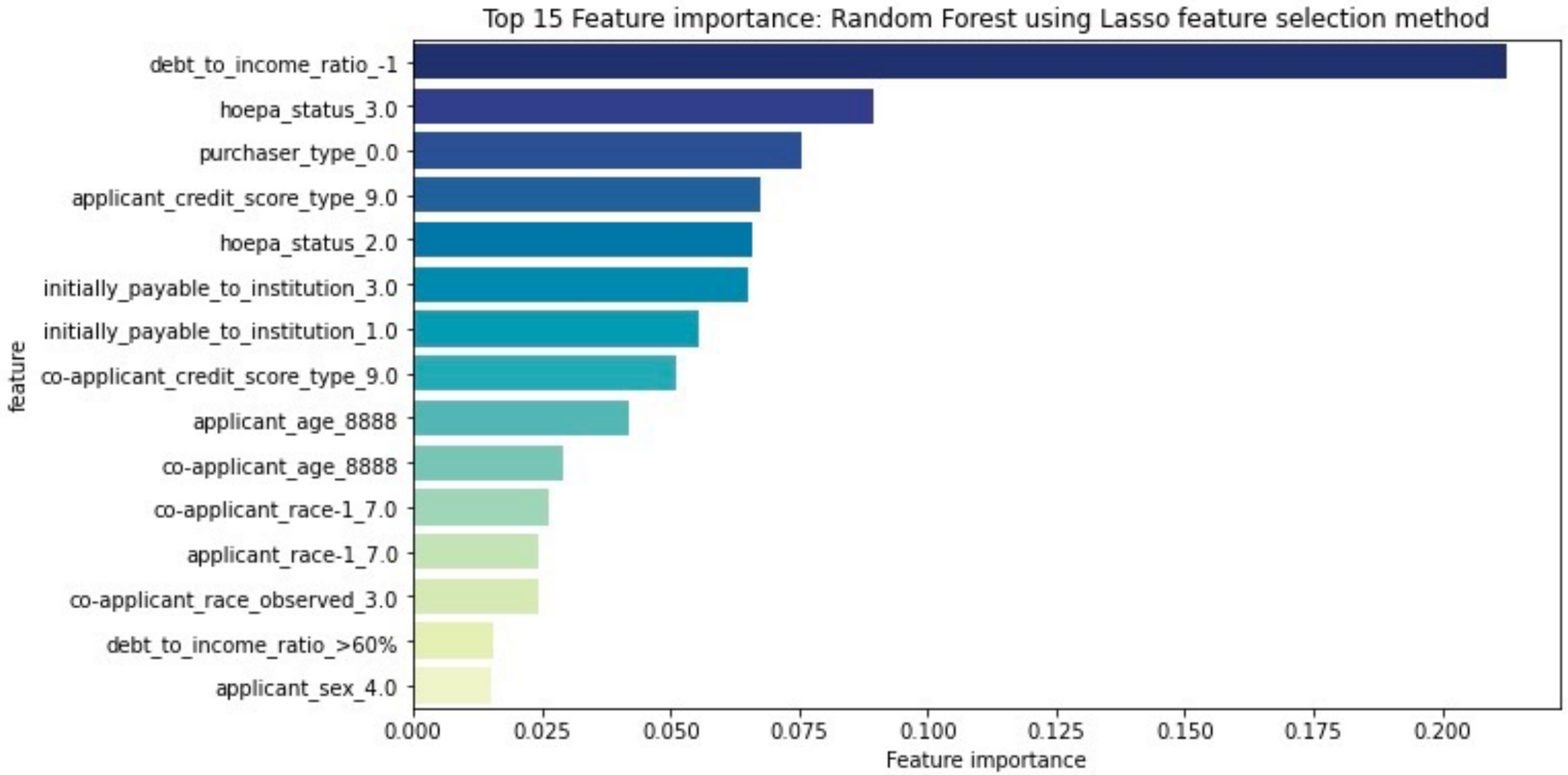
- Optimal : Random Forest Model using Lasso feature selection approach
- Generally, Random Forest is slightly better than Logistical Regression model.
- Lasso vs Pearson: Pearson Correlation only address the relationship between a single variable and the label, while Lasso method takes multivariable into consideration.
- PCA: When reducing dimensionality, the collinearity of variables makes the number of principals smaller, which do harm to the accuracy.

Content

1. Introduction
2. Data Prepossessing
3. Data visualization
4. Feature Engineering
5. Model Selection
6. Further Analysis

Further Analysis

Feature Importance of Random Forest with Lasso Feature Selection Method



Further Analysis

1 Fairness Analysis

Protected Attribute(X)	Accuracy from Optimal Model	
	With X	Without X
Sex	91.02%	91.02%
Race and Ethnicity	91.02%	90.99%

- We eliminate features contain gender information and race information to check the accuracy differences separately.
- **Conclusion:**

  Gender and Sex : Fair

  Race and Ethnicity: Nearly Fair

2 Limitations and Potential Improvements

- Computational Efficiency:

  Trade-off between improvements of accuracy and Runtime — 3% improvements with hours runtime

- Completeness of Data:

  “Not Applicable Data”

- Feature Selection Method:

  Better feature subset may be used—apply backward/forward algorithms with relative importance measure

Thank you!