Loan Prediction
Based on Customer Behavior

2021.11.29

Contents

1. Introduction

Our aim is to predict who possible defaulters are for the consumer loans product and find their main features. We have collected data about historic customer behavior based on what they have observed from Kaggle [1]. Hence when financial institutions have new clients, they can predict who is high-risk customer and who is not to reduce the risk of financial loss.

Our problems:

- Predict who possible defaulters are for the consumer loans product with models
- Find the features of customers who are more likely to default on a loan

Steps of Analysis:

- Data Preparation
- Exploratory Data Analysis
- Decision Trees
- Regression
- Neural networks
- Random forest
- Model assessment
- Conclusions and Suggestions
- References

2. Data Preparation

2.1 Data Description

Our dataset has 253000 records and 12 attributes of customers. Our target column is *Risk_Flag*.

Table 2.1 Data Description

Column	Data Type	Description	Example
Income	Interval	Income of the user	7060569
Age	Interval	Age of the user	48
Experience	Interval	Professional experience of the user in years	10
Married/Single	Nominal	Whether married or single	single, married
House_Ownership	Nominal	Owned or rented or neither	rented, owned, norent_noown
Car_Ownership	Nominal	Does the person own a car	yes, no
Profession	Nominal	Profession	Physician…
CITY	Nominal	City of residence	Bhopal…
STATE	Nominal	State of residence	Mizoram…
CURRENT_JOB_YRS	Interval	Years of experience in the current job	3
CURRENT_HOUSE_YRS	Interval	Number of years in the current residence	2
Risk_Flag(Target Role)	Binary	Defaulted on a loan	0— not defaulted 1— defaulted

2.2 Missing Value and Outlier Detection

We check the box plots of numerical variables like *Age, Experience, Current_job_yrs, Income and Current_house_yrs*, and we notice that there are no outliers. Besides, we also check null values, there is no missing value in this dataset.

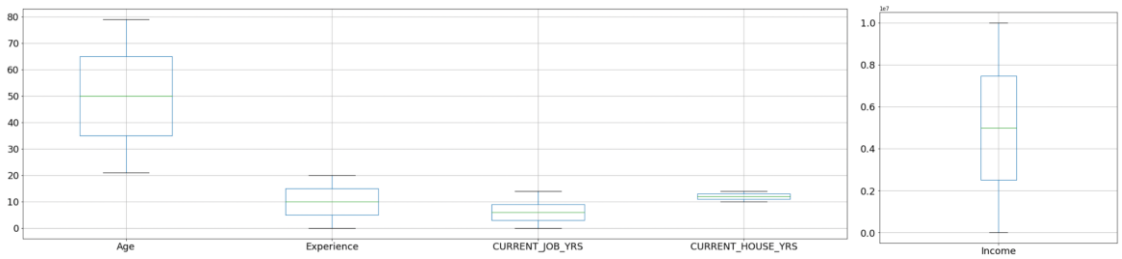


Figure 2.1 Box Plots

2.3 Data Transforming

We replace categorical variables (*'Married/Single', 'House_Ownership', 'Car_Ownership', 'Profession', 'CITY', 'STATE'*) to dummy, which will bring convenience to our model building. Besides, we regularize *Income* which has a large scale (from 536570 to 7936020) and may bring errors during analysis.

3. Exploratory Data Analysis

3.1 Density of Numerical Variables

We simply find that the majority of people who take loans has 2.5-6 years of experience in the current job. While for other attributes, the number of people is almost evenly distributed.

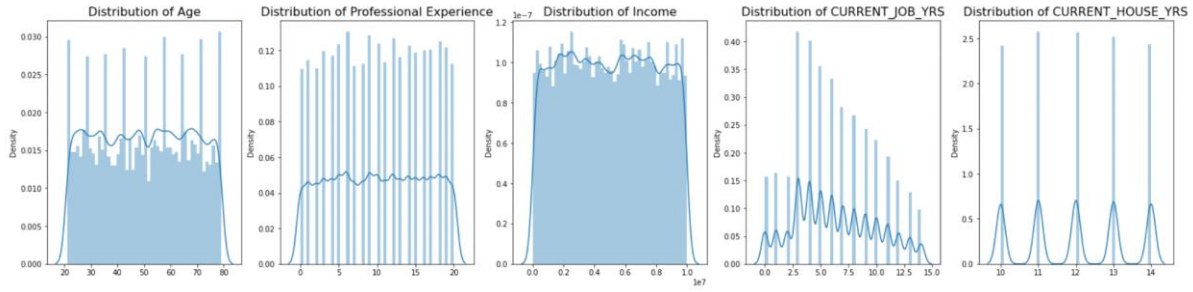


Figure 3.1 Density Plots

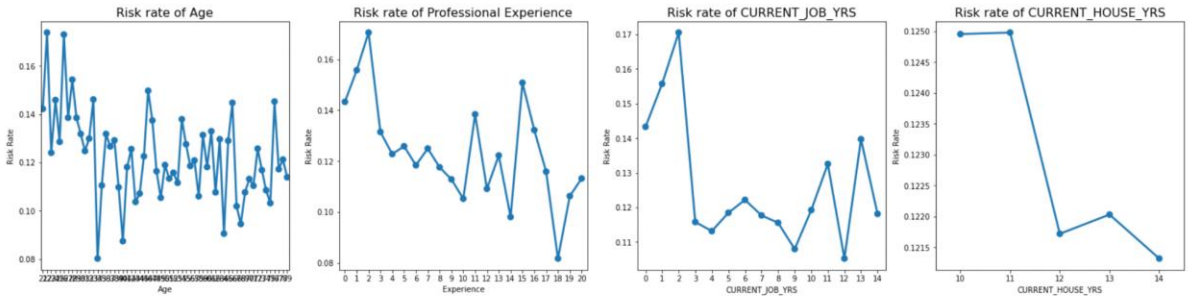


Figure 3.2 Risk Rate of Categorical Variables

For different Age, Professional Experience, Current join years and current house year, we calculate their risk rate. In general, young people aged under 30, people with less than 5 year professional experience or work less than 2 years in their current job or living less than 11 years in their current house are more likely to default on a loan.

3.2 Correlation analysis

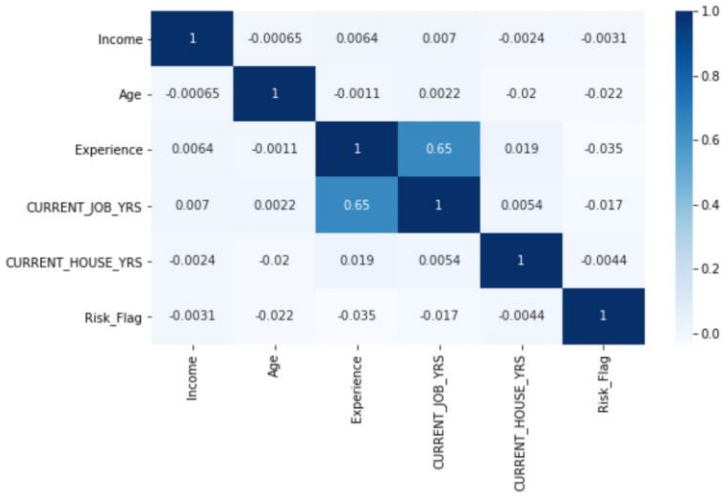


Figure 3.3 Correlation Matrix

From the correlation matrix, we can find all the numerical variables are not highly correlated.

3.3 Distribution of Categorical Variables

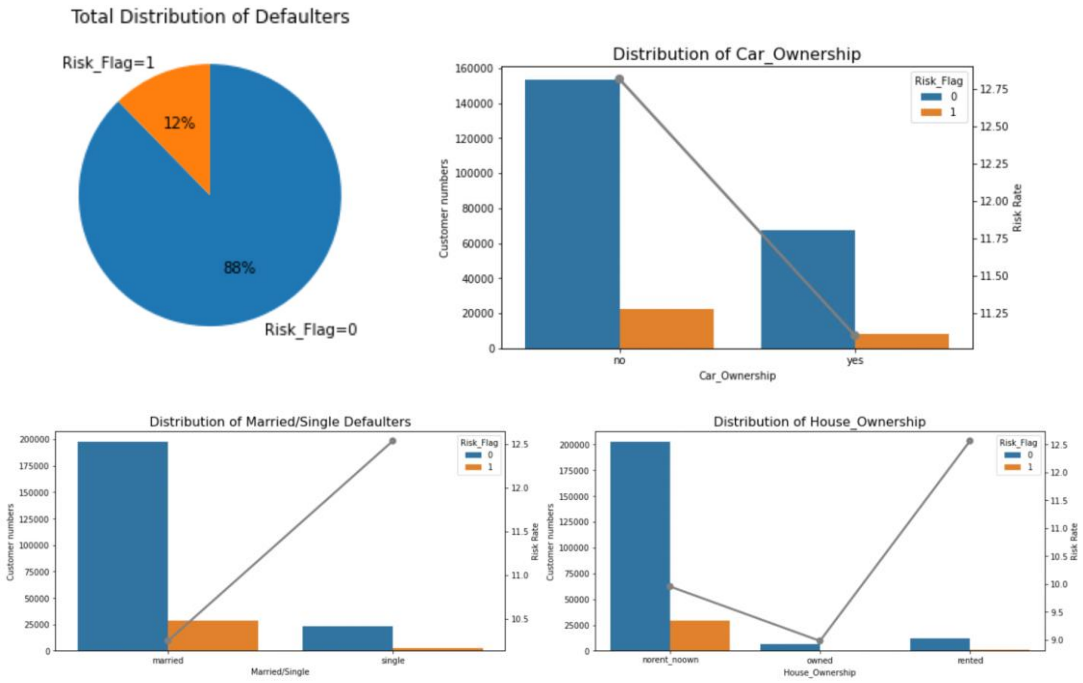


Figure 3.4 Distribution of Categorical Variables

Only about 12% of borrowers defaulted on a loan, and people who are single or without car ownership or living in a rented house are more likely to default on a loan than those who don't. That's quite reasonable as those people do not have to shoulder more responsibility than others.

3.4 Income and Risk Rate Rank

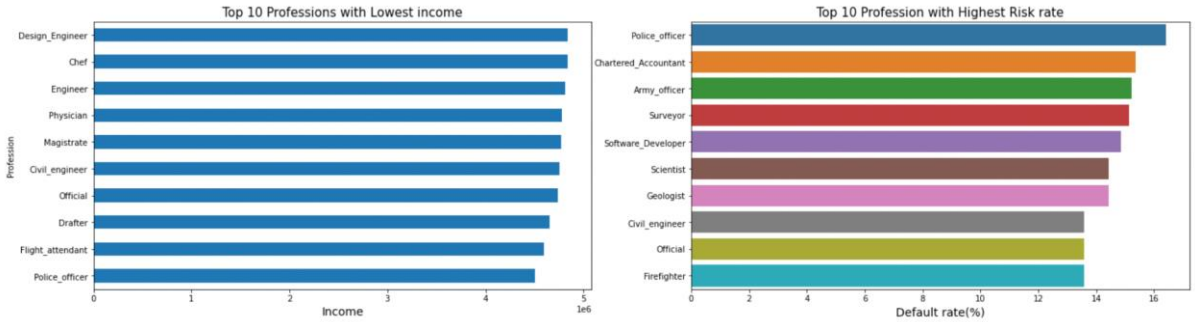


Figure 3.5

There is an interesting finding that Police officer is the top 1 profession with the highest risk rate as well as lowest income. Besides, the same patterns can be also found for Official and Cival_engineer. While, customers whose occupations are engineer, designer, dentist and analyst have less risk for default.

3.5 Risk Rate Rank by State

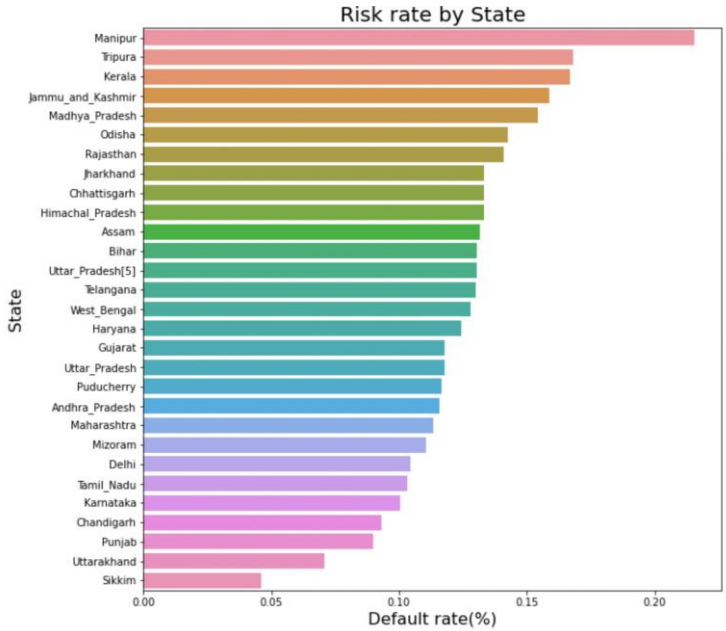


Figure 3.6 Risk Rate by State

People in states like Manipur, Tripura and Kerala have higher risk rates to default on a loan.

3.6 Conclusion

People without cars, living in a rented house or without Legal spouse are more likely to default on a loan. Besides, people with less than 2 years experience in their current job/whole profession have a higher risk rate than others.

Based on our EDA finding, here are some suggestions for companies.

People with 2.5-6 years experience in their current job are our main customers and they also have lower default rate, some activities can be held to attract them like sending messages.

For our main customers with lower default rates ('Industrial_Engineer', 'Chemical_engineer','Web_designer', 'Mechanical_engineer', 'Drafter'), we can put on some marketing events like put ads in their working buildings or neighbor subway stations.

4. Logistic Regression

4.1 Model Introduction

Logistic regression, also known as Logistic regression analysis, is a generalized linear regression analysis model, which is often used in data mining, automatic diagnosis of diseases, economic forecasting and other fields. For example, to explore the risk factors causing the disease, and predict the probability of disease occurrence according to the risk factors. Taking gastric cancer disease analysis as an example, two groups of people were selected, one group was gastric cancer group, the other group was non-gastric cancer group, the two groups of people must have different physical signs and lifestyle, etc. Therefore, the dependent variable is whether gastric cancer, and the value is "yes" or "no", and the independent variable can include many, such as age, gender, eating habits, helicobacter pylori infection, etc. Independent variables can be either continuous or classified. Then, the weight of independent variables can be obtained through logistic regression analysis, so as to roughly understand which factors are risk factors of gastric cancer. This weight can also be used to predict a person's likelihood of developing cancer based on risk factors.

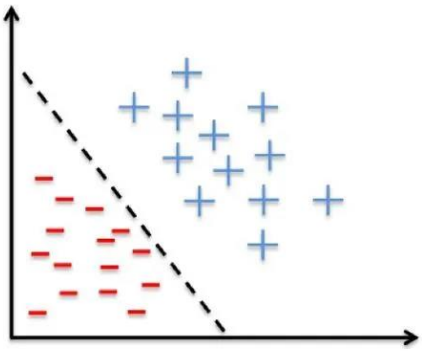


Figure 4.1 Logistic Regression Classification

Logistic regression is different from linear regression. The result of the linear regression is numeric. The prediction result can be any figure. But logistic regression is more like a classifier. It can classify the result of linear regression to a classification result. The picture below shows its relationship between the linear regression.

Table 4.1 Operation Process of Logistic Regression

Sample Input				sigmoid	Logistic regression result		Prediction result	Real result
12.3	20.0	16.0			0.4		B	A
9.4	21.1	7.2			0.68		A	B
34.4	18.7	8.1	* W =	→	0.41	→	B	A
10.2	16.0	12.5			0.55		B	B
5.6	10.0	6.3			0.71		A	A

4.2 Model Loss and Optimization

The loss of logistic regression is logarithmic likelihood loss, with the help of the log thought. The following figure is the calculation function of classified loss, which y means the true value.

$$cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if y=1} \\ -log(1 - h_\theta(x)) & \text{if y=0} \end{cases}$$

Figure 4.2 Operation Function of Logistic Regression

Gradient descent optimization algorithm is used to reduce the value of the loss function. In this way, the weight parameters of the corresponding algorithm before logistic regression are updated to improve the probability that originally belongs to the 1 category and reduce the probability that originally belongs to the 0 category.

4.3 Classify Evaluation Results

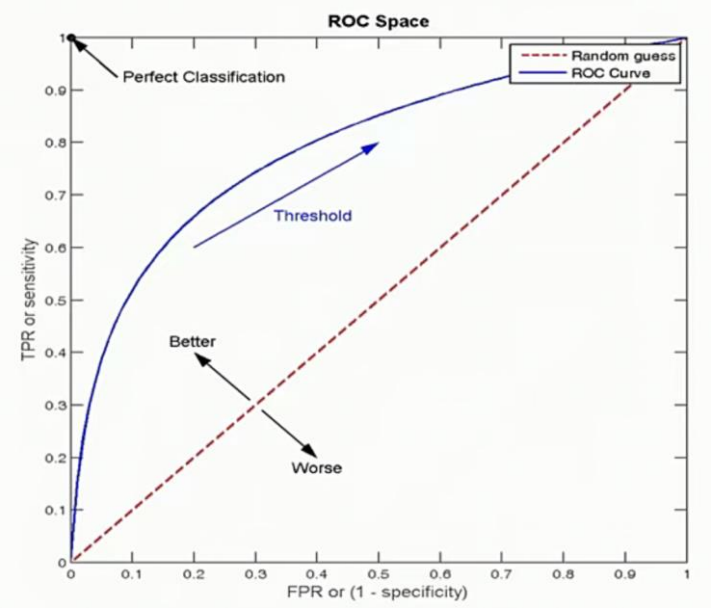


Figure 4.3 ROC Curve and AUC

Table 4.2 True and Prediction

	Prediction		
TRUE		1	0
	1	TP	FN
	0	FP	TN

The picture above shows the principles in the ROC curve. TPR = TP/(TP + FN); FPR = FP/(FP+TN). The horizontal axis of the ROC curve is FPRate and the vertical axis is TPRate. When the two are equal, it means that the probability of the classifier predicting to be 1 is equal for the samples no matter whether the category is really 1 or 0, and the AUC is 0.5.

4.4 Python Result of Model

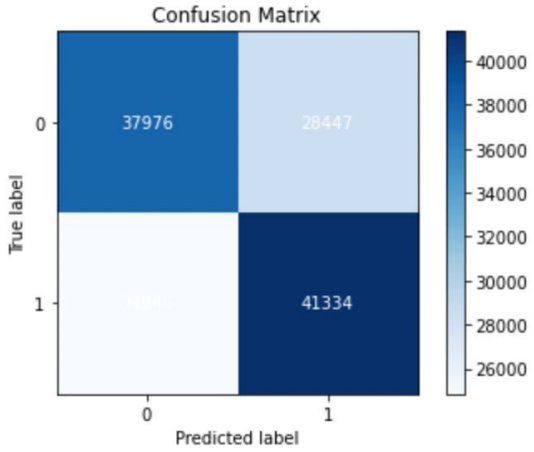


Figure 4.4 Confusion Matrix

The table above is the confusion matrix of the logistics regression.

	precision	recall	f1-score	support
not default	0.60	0.57	0.59	66423
default	0.59	0.62	0.61	66180
accuracy			0.60	132603
macro avg	0.60	0.60	0.60	132603
weighted avg	0.60	0.60	0.60	132603

Figure 4.5 Evaluation Score of Model

Table below shows the related rates of the prediction. Prediction scores for not default and default are 0.60 and 0.59. Recall scores for not default and default are 0.59 and 0.61, which shows the prediction of the model is not good because of the low score.

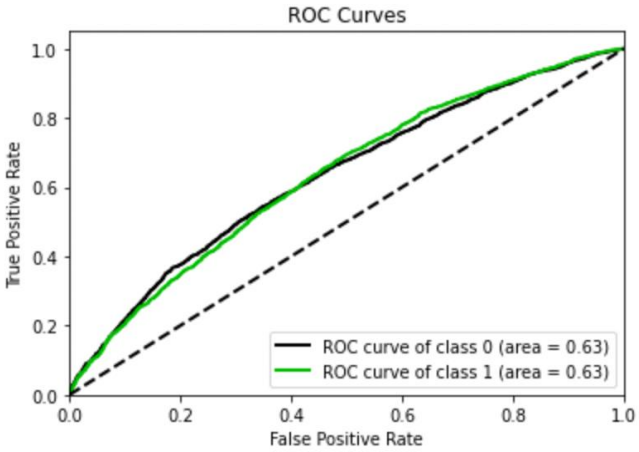


Figure 4.6 ROC Curve

Also, the ROC curve above and AUC value which is only 0.63 shows the prediction of the model is not good.

4.5 Logistic Regression Conclusion

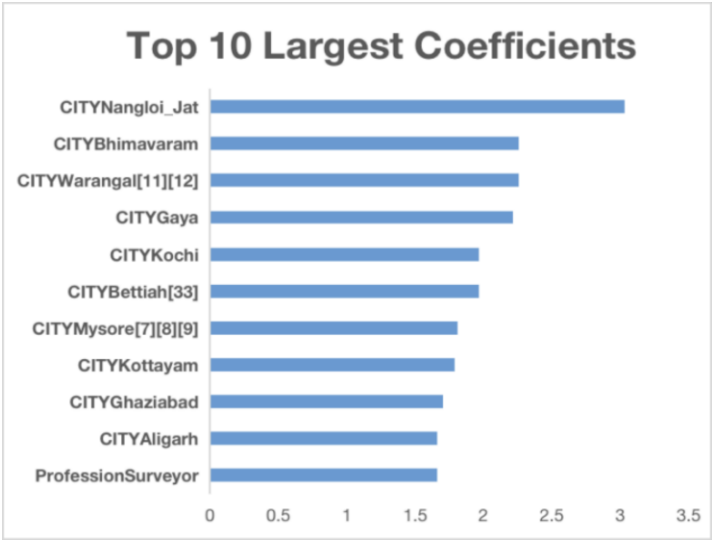


Figure 4.7 Ordered Coefficients of Model

From the sorted result of the coefficients of variables above, we can find People from cities like Nangloi_Jat, Bhimavaram and Warangal[11][12] are more likely to default since they have higher coefficients in the regression model. The prediction of the model is not good which may need further improvement.

5. Decision Tree

5.1 Decision tree structure

We use python to build a decision tree and we can clearly see the structure of the decision tree from the code and plot.

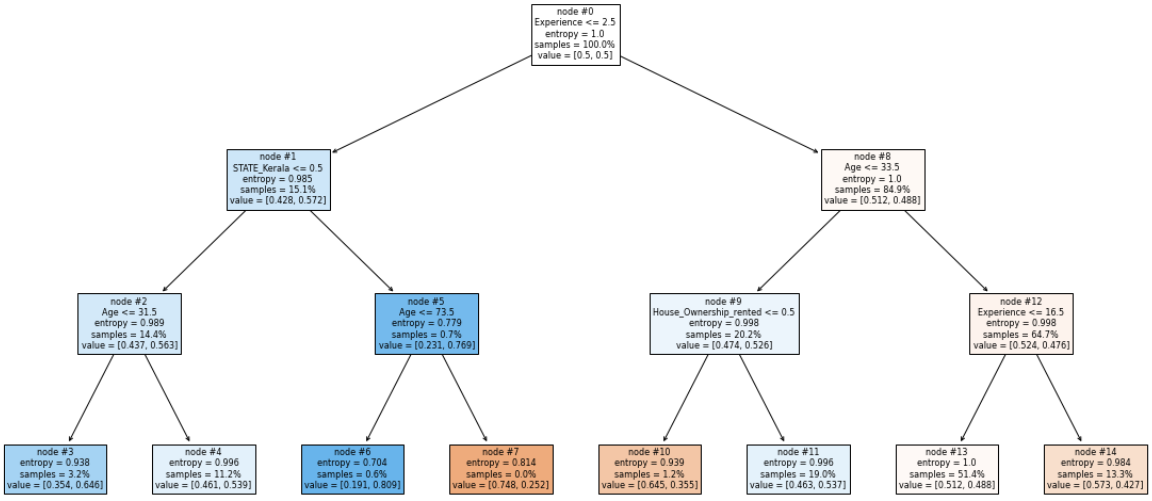


Figure 5.1 Decision Tree

As can be seen from the plot, the criterion of this decision tree is entropy, and we define the depth of the decision tree as 3. The structure of our decision tree is clearly shown in the above plot, the bottom of the plot shows the leaves of the tree, which has a number of 8. In each node, we can see the detailed parameters of each step. I will take the root decision node as an example as follows.

5.2 Decision trees root decision node

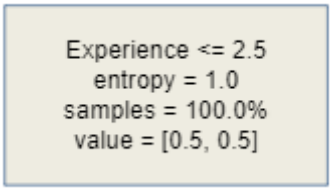


Figure 5.2 root decision node

In the root decision node, the first line shows the judging criteria of this decision node is based on experience. When the value of experience is lower than or equal to 2.5, the node produce the results of true and get the input separated into the right section.

The second line shows the result of entropy, and the format of entropy is as follows:

$$Entropy(s) = \sum_{i=1}^{c} -p_i log_2 p_i$$

The following line shows the percentage of the input samples, in terms of this node is the root decision node, the value of the percentage of sample is 100%.

And the last line shows the distribution of TRUE and FALSE in the result. In this node, the value is [0.5,0.5]

5.3 Decision trees confusion matrix & ROC

Then we use the confusion matrix and ROC curves of the optimized model, which are close to perfect.

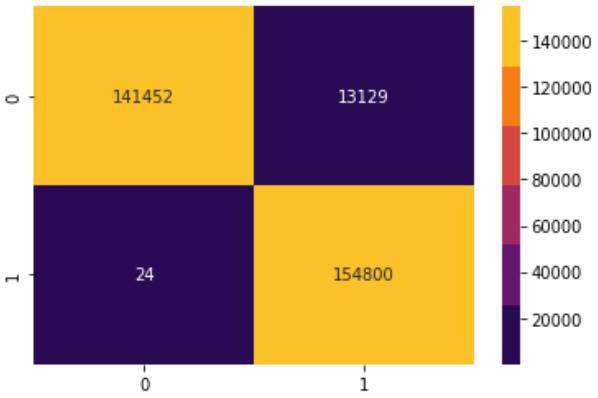


Figure 5.3 Confusion Matrix

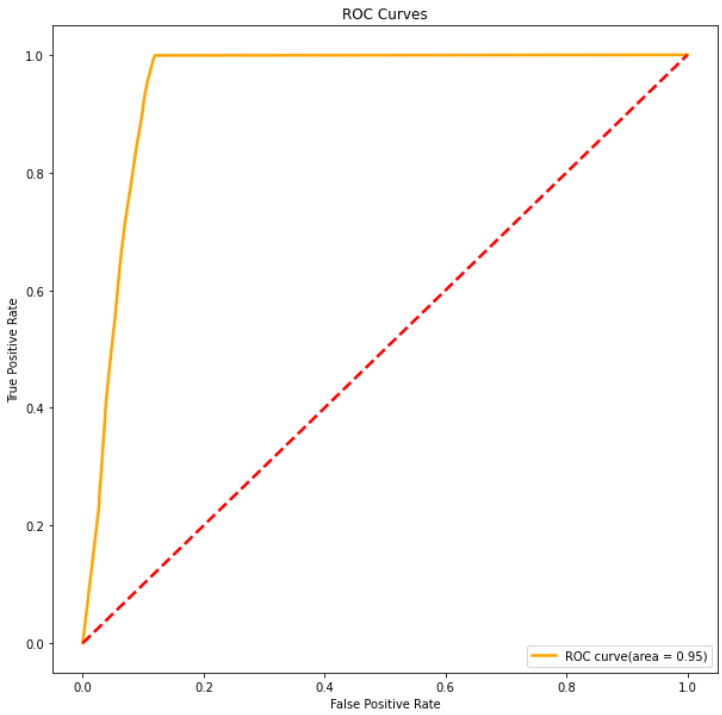


Figure 5.4 ROC Curves

5.4 Conclusion

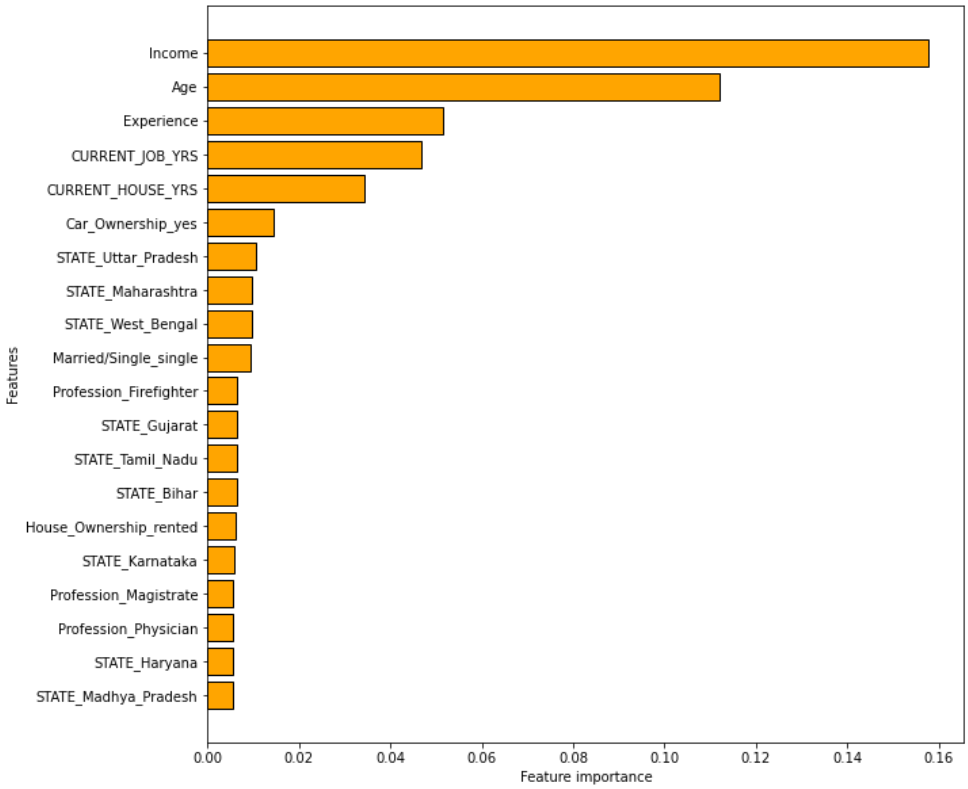


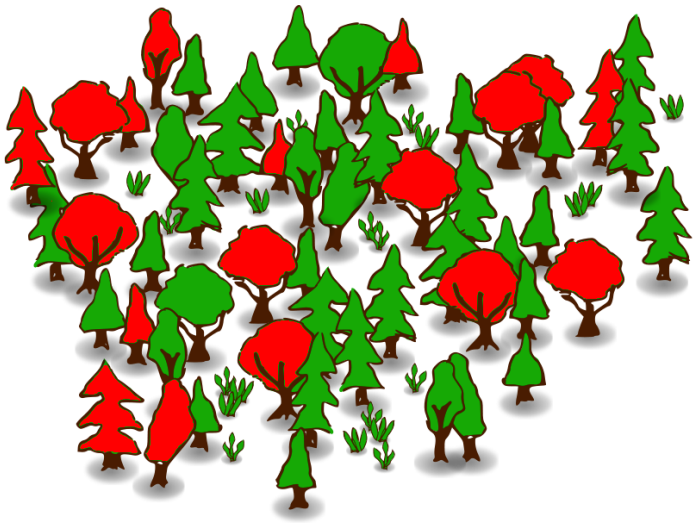
Figure 5.5 Conclusion

From the results of the importance of decision tree features, it can be seen that the top five most important influencing factors are: Income, Age, Experience, CURRENT_JOB_YRS and CURRENT_HOUSE_YRS.

So we can draw to the conclusion that, We can choose high consumption capacity people and older people as main customers for the reason that the risk of these groups is lower; According to the frequency of the label of people in the decision tree, we can use targeted advertising to decrease the cost.

6. Random Forest

6.1 What is random forest



Random forest is a classifier that contains multiple decision trees. For an input sample, different trees in the random forest will yield different classification results based on different preferences. The random forest combines the voting results of all trees and finally the result with the most votes is the final result output.

Random forest has the following advantages:

1. the ultra-high accuracy of random forests compared to other algorithms such as logistic regression and decision trees.

2. efficiency in handling large data sets.

3. the ability to handle input samples with high-dimensional features without dimensionality reduction.

4. the ability to evaluate the importance of individual features on the classification problem.

5. the ability to obtain an unbiased estimate of the internal generation error in the generation process.

6. the problem of missing values can be handled well.

3) Every tree grows to the maximum extent, and there is no pruning process.

The generation rule for each tree in a random forest is based on the following steps.

1) For each tree in the random forest, randomly and with put-back draw N training samples from the training set (assuming the size of the training set is N) (i.e., bootstrap sample method), and use the obtained samples as the training set for this tree.

2) Assuming that each sample has M feature dimensions, the random forest will specify a constant m (which value is much smaller than M), and a subset of m features will be randomly selected from the sample M features, and the best one will be selected from these m features each time the tree splits.

3) Every tree in a random forest grows to the maximum possible extent and we do not perform the pruning process.

6.2 Why use random forest

In order to predict and classify whether to lend or not according to the customer's behavior more accurately, we introduced an optimized version compared to the decision tree, namely random forest.

Random forest uses a voting mechanism of multiple decision trees to improve decision trees.

6.3 Random forest performance

Here we divide the original data set into training data set and validation data set.

From the model scores table, we can see that ROC AUC Score of the default model is greater than 0.9, which means that the prediction performance of random forest is quite perfect.

Table 6.1 Model Scores

Model 1 score:	0.8431410504581236
Model 1 Precision:	0.8811561666900519
Model 1 Recall:	0.7985757884028484
Model 1 F1 score:	0.8378360349543059
Model 1 Accuracy:	0.8431410504581236

Model 1 ROC AUC Score:	0.9310391786783494

Similarly, let us look at the confusion matrix and ROC curves of the optimized model, it is also close to perfect.

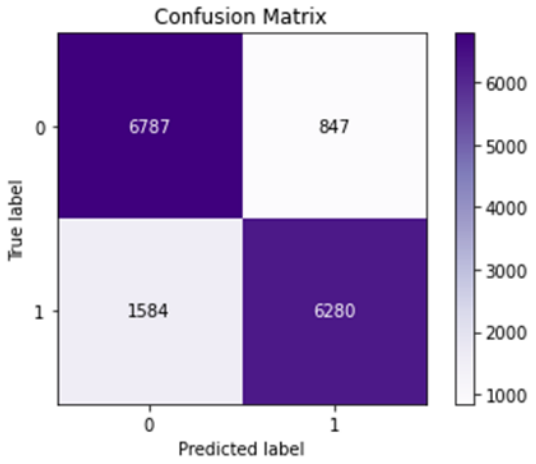


Figure 6.1 Confusion Matrix

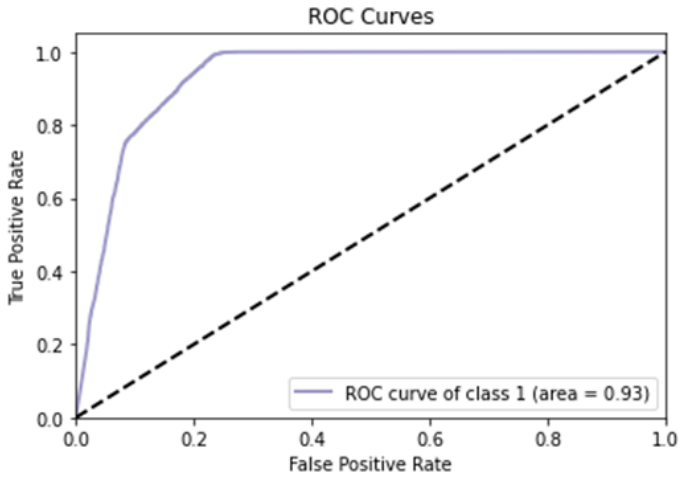


Figure 6.2 ROC Curves

6.4 Conclusion

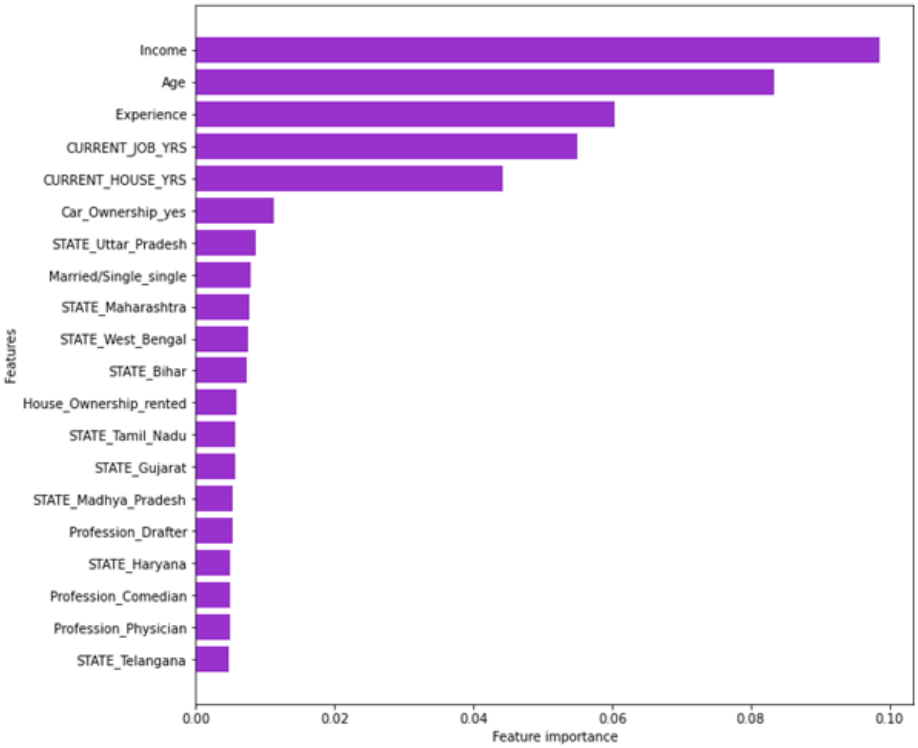


Figure 6.3 Feature Importance

From the results of the importance of random forest features, it can be seen that the top five most important influencing factors are: Income, Age, Experience, CURRENT_JOB_YRS and CURRENT_HOUSE_YRS. The impact of these five factors on whether there will be loan risks exceeds 38%.

This seems to be intuitive, that is, the higher the income, the older the age, and the more experienced people are less likely to default.

In this regard, we should be cautious in lending to people with low incomes, and we should strengthen responsible publicity for young people, popularize the benefits of timely loan repayment, and foster good credit among them.

7. Artificial Neural Network

Over the past years, there has been a growing consensus that computing based on models inspired by our understanding of the structure and function of biological neural networks could be the key to machine intelligence. Artificial Neural Networks is the name of the new field.

7.1 Introduction of artificial neural networks

Artificial neural networks (ANNs), sometimes known as neural networks (NNs), are computer systems that are modeled after biological neural networks. ANNs are becoming increasingly popular in many fields of science, and they are also being used to handle a variety of industrial and civil challenges. The design of ANN is motivated by analogy with the brain. Human brain consists of neural cells. In comparison to the entire brain, the function of neurons is shockingly simple; each biological neuron consists of a cell body, a collection of dendrites that bring information into the cell, and an axon that transmits information out of the cell. (See Figure 7.1).

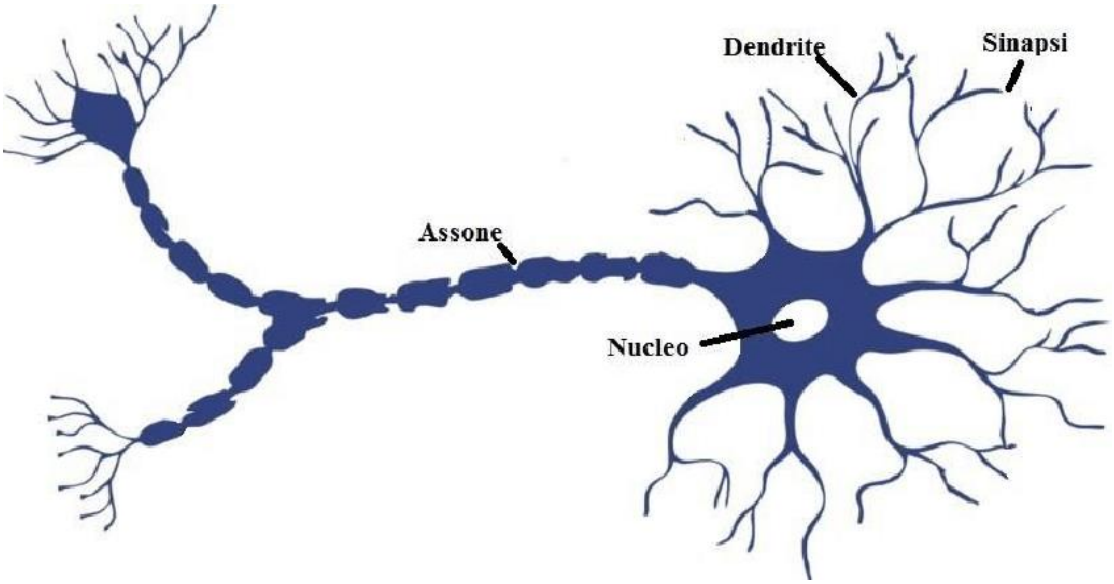


Figure 7.1 Neural cell

ANN is a network of interconnected artificial neurons. In a simple term, an artificial neuron is a processor which generates a response to some set of weighted inputs. This process can be approximated by model shown in Figure 7.2

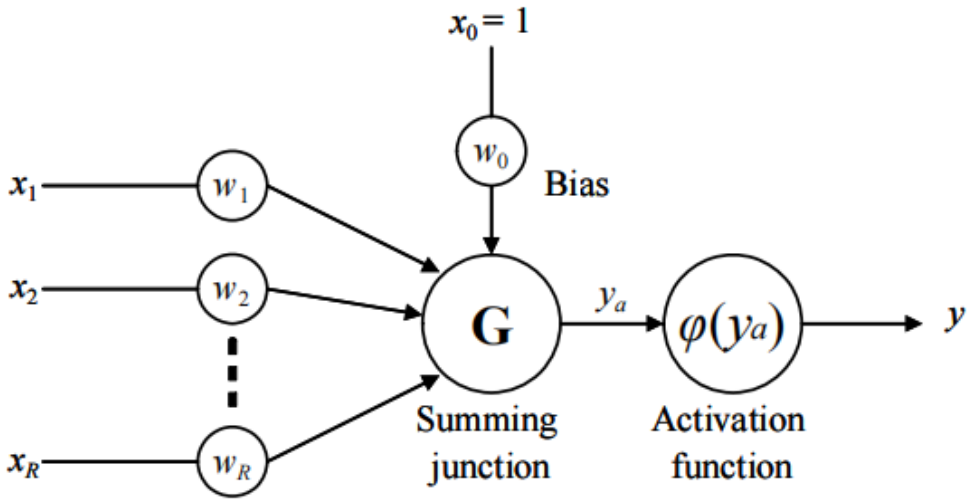


Figure 7.2 Artificial neuron

Here are the fundamental components:

- A collection of interconnecting links, each with its own weight. A signal x_i coupled to a neuron, for example, is multiplied by a weight w_i.
- An adder that combines the weighted input signals into a single scalar value known as activation potential.
- The activation function processes the activation potential and generates the neuron's output.

Artificial neuron in Figure 7.2 also includes an externally bias $w_0 x_0$. The input to the activation function is influenced by the bias.

$$y_a = \sum_{i=0}^{R} w_i \, x_i$$

$$y = \varphi(y_a)$$

Activation function, which is denoted by $\varphi(.)$, can be defined in a variety of ways, and its definition has a significant impact on the overall response of the ANN. In Figure 7.3, there are graphs of the most common activation functions.

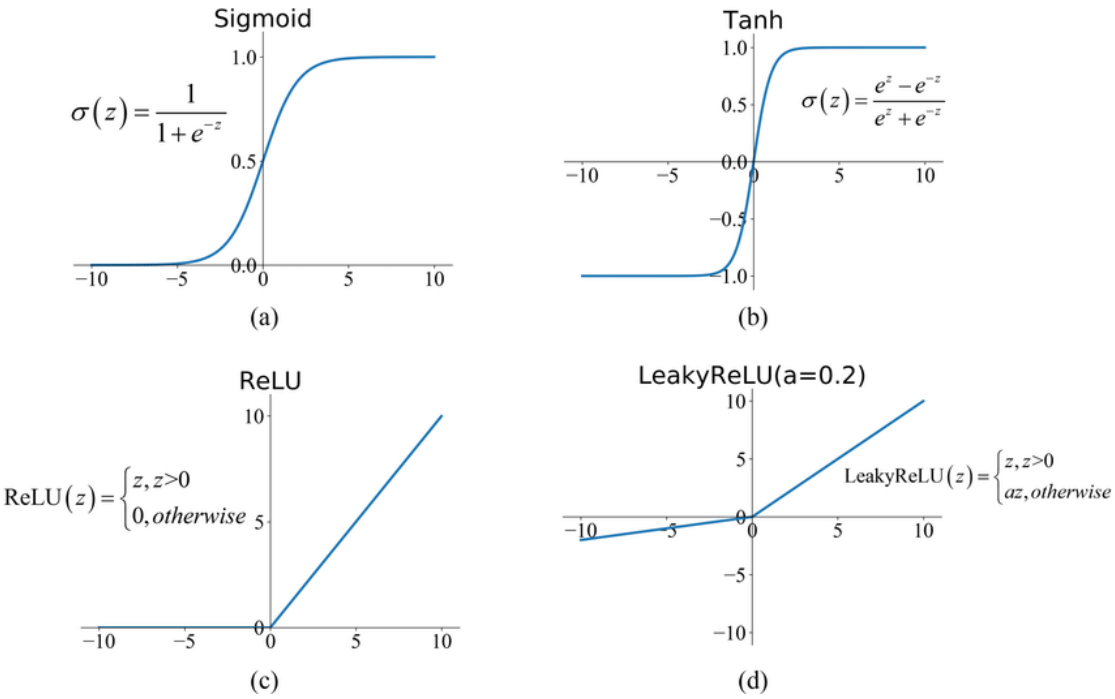


Figure 7.3 Common activation functions

The neuron's architecture (number of inputs, adder definition, and activation function) is essentially constant, while the other parameters (particularly weights and biases) are modified through learning. One neuron, however, can only answer easy problems. As a result, it is suggested that the neurons be connected to form a neural network.

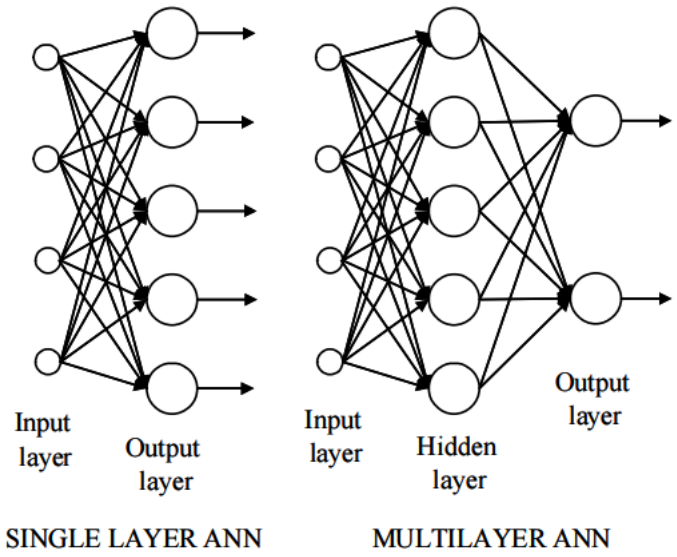


Figure 7.4 Feedforward ANNs

ANN consists of one or more neurons connected into one or more layers. A layer is usually made up of neurons that are not coupled to one another in any way (there are some exceptions, of course). Examples of ANN possible topologies are shown in Figure 7.5. The topology of the ANN depends on

a task to be solved. Tasks like function approximation or pattern recognition can be solved using feedforward ANN with one or two hidden layers. The weights and biases of each neuron in an ANN must be fine-tuned in order for it to function effectively. This process is called learning.

7.2 Modelling process and model results

First the data pre-processing step, which focuses on encoding and standardization.

Then create an ANN model with 2 hidden layers and 64 hidden neurons in each layer. Main parameters in an ANN model are:

Epochs: One Epoch is when an entire dataset is passed forward and backward through the neural network only once. A total of 100 epochs are employed in the model.

Batch Size: Total number of training examples present in a single batch.

Optimizer: Optimizers are algorithms or methods used to minimize an error function (loss function) or to maximize the efficiency of production. Adam optimization, a stochastic gradient descent method based on adaptive estimate of first- and second-order moments, is chosen.

The ANN model's accuracies are shown in Figure 7.5.

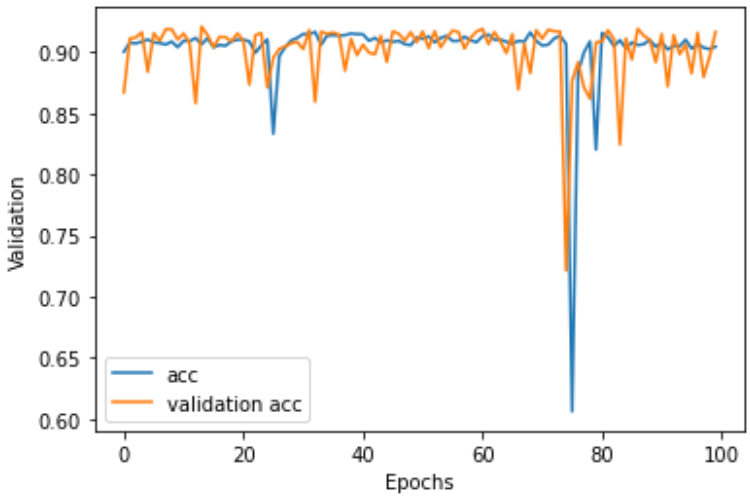


Figure 7.5

It can be found that the results are not ideal since it is not stable. So training process is necessary.

The ANN training process aims to build the complex nonlinear relationships between the features and the demand by minimizing the mean squared prediction error.

Using the grid search algorithm tuning the parameter “batch_size”, “epochs”, “optimizer”. After tuning a set of best parameters could be got. They are 'batch_size': 64, 'epochs': 150, 'optimizer': 'adam'.

The accuracies use best parameters shown in Figure 7.6.

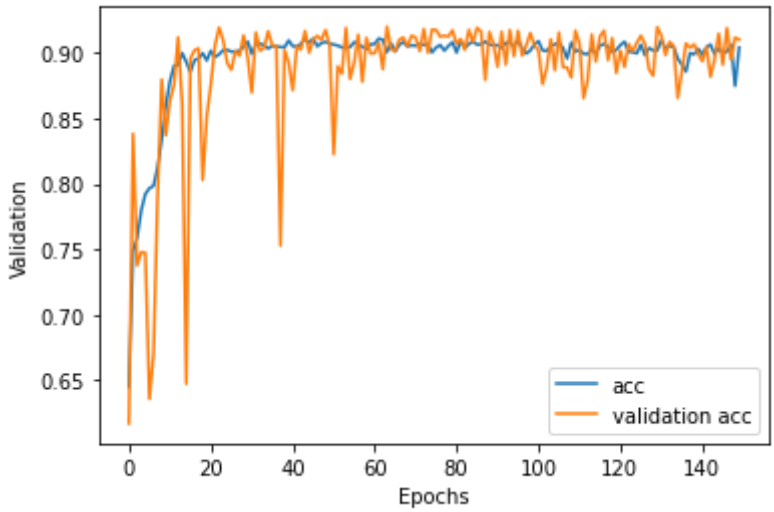


Figure 7.6

The accuracies after tuning parameter is better than before.

A figure of the model roc curve using the best parameters is below.

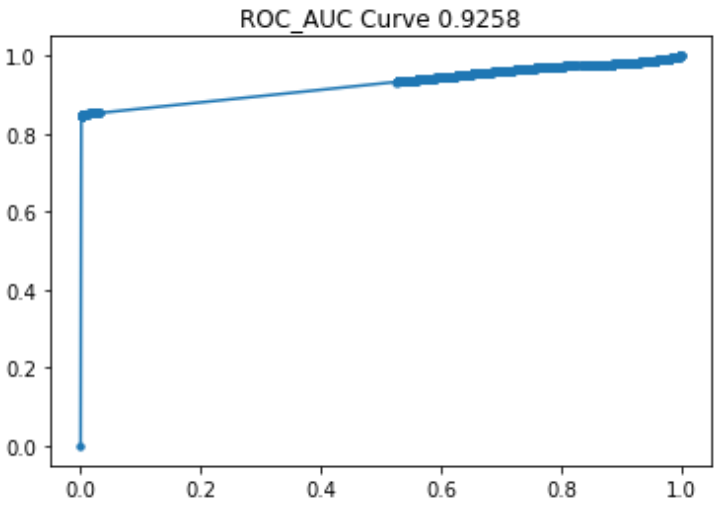


Figure 7.7

8. Conclusion

8.1 Model Comparison

Model Comparison (ROC AUC Score)			
Logistic Model	Decision Tree	Random Forest	ANN
0.63	0.95	0.93	0.93

Comparing the ROC AUC Score of the four models, we finally believe that the decision tree is the best model. So we can use this model to predict who are possible defaulters and reduce the default rate to cut loss.

8.2 Suggestion

Based on the results of the model, we draw the following conclusions:

- People with 2.5-6 years experience in their current job are our main customers and they also have lower default rate, some activities can be held to attract them like sending messages.
- For our main customers with lower default rates like 'Web_designer', 'Drafter' 'Industrial/Chemical/Mechanical_Engineer', we can put on some marketing events like put ads in their working buildings or neighbor subway stations.
- People from cities like Nangloi_Jat, Bhimavaram and Warangal[11][12] are more likely to default since they have higher coefficients in the regression model.
- We should be cautious in lending to people with low incomes, and we should strengthen responsible publicity for young people, popularize the benefits of timely loan repayment, and foster good credit among them.
- Comparing logistic regression, decision tree, random forest and ANN, we compare their ROC AUC score, and finally we think random forest and ANN are the best models. Here, we can use

them to reduce the loan default rate to decrease loss.

- We can choose high consumption capacity people and older people as main customers for the reason that the risk of these groups is lower; According to the frequency of the label of prople in the decision tree, we can use targeted advertising to decrease the cost.

9. References

1. Loan Prediction Based on Customer Behavior. (2021). Retrieved 28 November 2021, from https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior

2. Blog.csdn.net. 2021. *决策树的 Python 实现（含代码）_CDA 数据分析师-CSDN 博客_决策树 python*. [online] Available at: <https://blog.csdn.net/yoggieCDA/article/details/92832367> [Accessed 28 November 2021].

3. Bayya, Y. (2009). ARTIFICIAL NEURAL NETWORKS.

4. Petr, D., Martin, M., & Ivan. (2013). Artificial Neural Network Promotion How to introduce Artificial Neural Networks to Students. 2013 INTERNATIONAL CONFERENCE ON PROCESS CONTROL (PC).

5. Team, K. (2021). Keras documentation: Adam. Retrieved 30 November 2021, from https://keras.io/api/optimizers/adam/