Risk Prediction

Paidy

Agenda

1. Data Quality Check
2. Data Distribution
3. Feature Collinearity
4. Risk Distribution
5. Xgboost Model
6. High-Risk Validate
7. Logistic Regression
8. CNN Model
9. Conclusion

01

Data Quality Check

Train Dataset Missing Rate

	Missing Values	% of Total Values	Data Type
MonthlyIncome	29731	19.8	float64
NumberOfDependents	3924	2.6	float64
id	0	0.0	int64
SeriousDlqin2yrs	0	0.0	int64
RevolvingUtilizationOfUnsecuredLines	0	0.0	float64
age	0	0.0	int64
NumberOfTime30-59DaysPastDueNotWorse	0	0.0	int64
DebtRatio	0	0.0	float64
NumberOfOpenCreditLinesAndLoans	0	0.0	int64
NumberOfTimes90DaysLate	0	0.0	int64
NumberRealEstateLoansOrLines	0	0.0	int64
NumberOfTime60-89DaysPastDueNotWorse	0	0.0	int64

Note:

1. There are missing values on the MonthlyIncome and NumberOfDependents fields

2. MonthlyIncome field has a relatively high missing rate compared with NumberOfDependents

02

Data Distribution

	count	mean	std	min	1%	5%	25%	50%	75%	95%	99%	max
id	150000.0	75000.500000	43301.414527	1.0	1500.99	7500.950000	37500.750000	75000.500000	112500.250000	142500.05	148500.010000	150000.0
SeriousDlqin2yrs	150000.0	0.066840	0.249746	0.0	0.00	0.000000	0.000000	0.000000	0.000000	1.00	1.000000	1.0
RevolvingUtilizationOfUnsecuredLines	150000.0	6.048438	249.755371	0.0	0.00	0.000000	0.029867	0.154181	0.559046	1.00	1.092956	50708.0
age	150000.0	52.295207	14.771866	0.0	24.00	29.000000	41.000000	52.000000	63.000000	78.00	87.000000	109.0
NumberOfTime30-59DaysPastDueNotWorse	150000.0	0.421033	4.192781	0.0	0.00	0.000000	0.000000	0.000000	0.000000	2.00	4.000000	98.0
DebtRatio	150000.0	353.005076	2037.818523	0.0	0.00	0.004329	0.175074	0.366508	0.868254	2449.00	4979.040000	329664.0
MonthlyIncome	120269.0	6670.221237	14384.674215	0.0	0.00	1300.000000	3400.000000	5400.000000	8249.000000	14587.60	25000.000000	3008750.0
NumberOfOpenCreditLinesAndLoans	150000.0	8.452760	5.145951	0.0	0.00	2.000000	5.000000	8.000000	11.000000	18.00	24.000000	58.0
NumberOfTimes90DaysLate	150000.0	0.265973	4.169304	0.0	0.00	0.000000	0.000000	0.000000	0.000000	1.00	3.000000	98.0
NumberRealEstateLoansOrLines	150000.0	1.018240	1.129771	0.0	0.00	0.000000	0.000000	1.000000	2.000000	3.00	4.000000	54.0
NumberOfTime60-89DaysPastDueNotWorse	150000.0	0.240387	4.155179	0.0	0.00	0.000000	0.000000	0.000000	0.000000	1.00	2.000000	98.0
NumberOfDependents	146076.0	0.757222	1.115086	0.0	0.00	0.000000	0.000000	0.000000	1.000000	3.00	4.000000	20.0

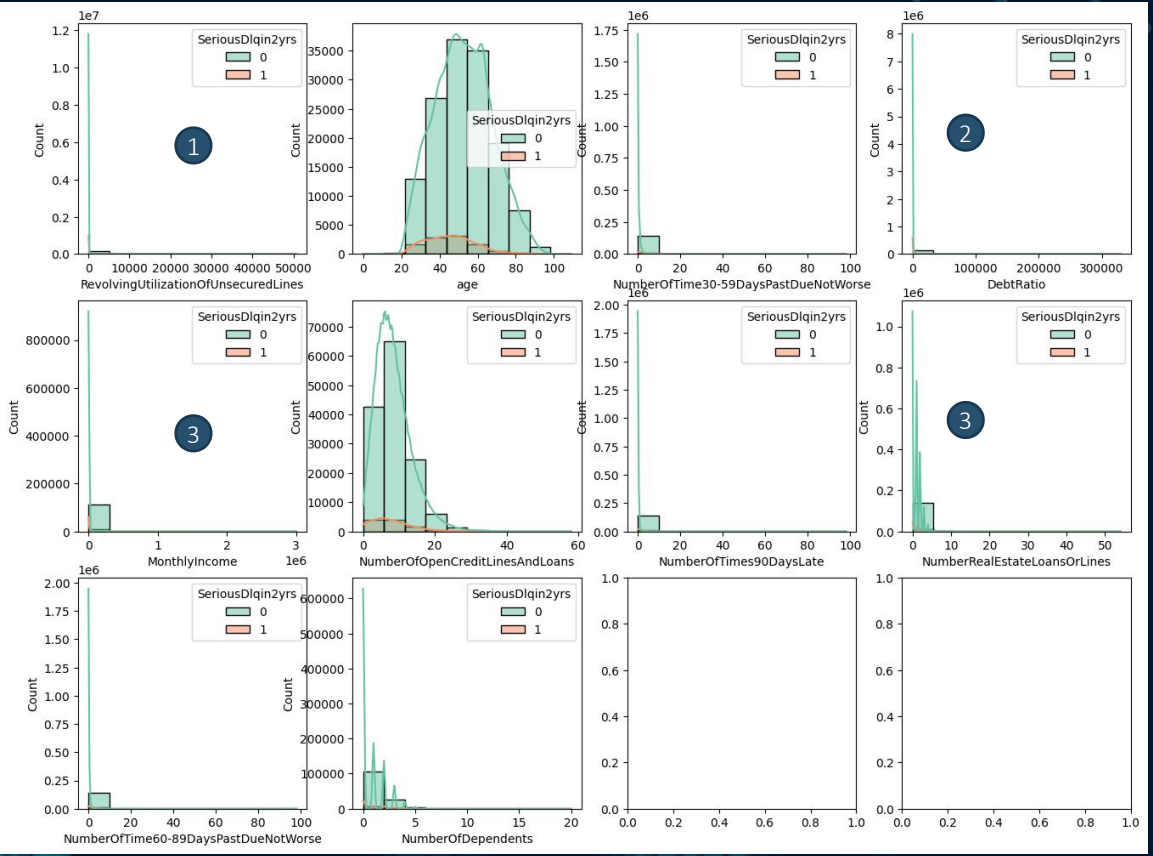
1 2 3 4

Note:

1. Age should not contain data with 0.
2. RevolvingUtilizationOfUnsecuredLines has some extreme values
3. Debt ratio has some outlier data
4. NumberOfDependents has some extreme / unreasonable values

02

Data Distribution



Note:

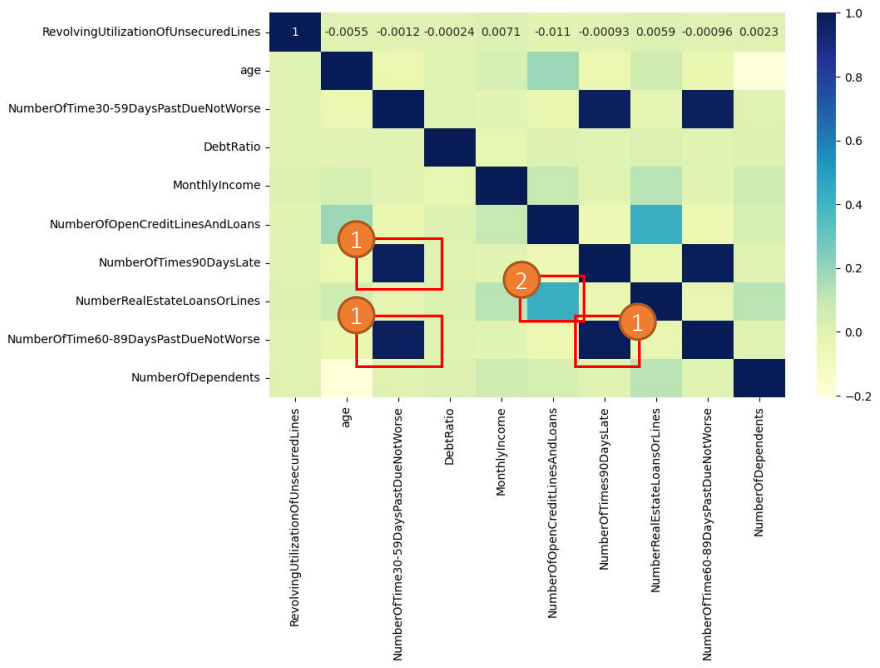
1. RevolvingUtilizationOfUnsecuredLines is mainly around 0 while there are some extreme values

2. DebtRatio field's distribution is strange. Most of the values is around 0 while there are some large values.

3. MonthlyIncome and NumberRealEstateLoansOrLines filed has some long-tail values on the right.

03

Feature Collinearity



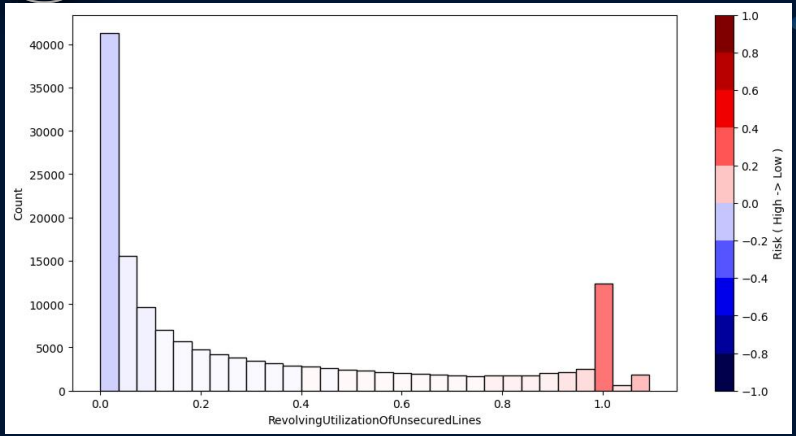
Note:

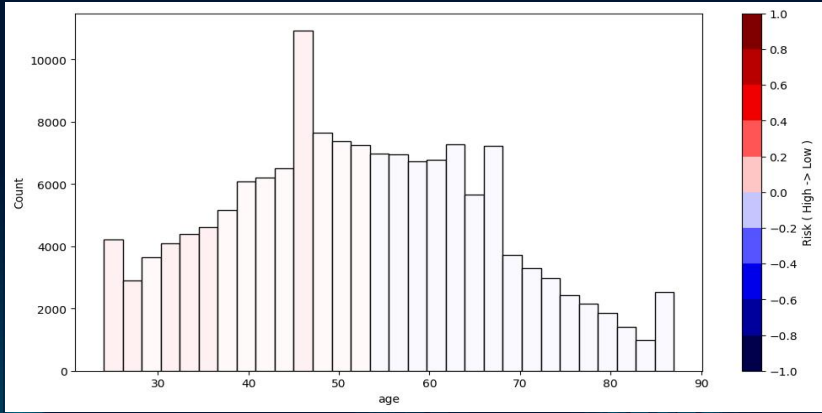
1. NumberOfTime60-89DaysPastDueNotWorse, NumberOfTimes90DaysLate and NumberOfTime60-89DaysPastDueNotWorse has high collinearity (Over 0.9).

2. NumberOfOpenCreditLinesAndLoans and NumberRealEstateLoansOrLines has relatively high relevance (over 0.45)

04

Risk Distribution





Note:

1. We checked the bad customer distribution on different features within maximum 30 bins using K-L distance or Cross-Entropy measurement.

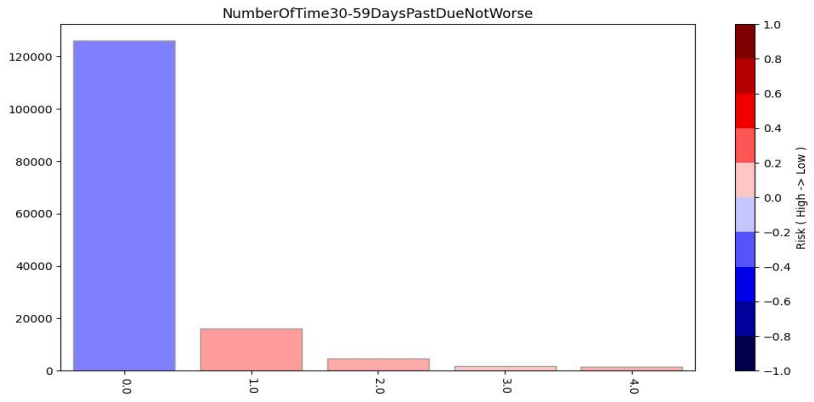
2. The red color means high risk while blue color means low risk

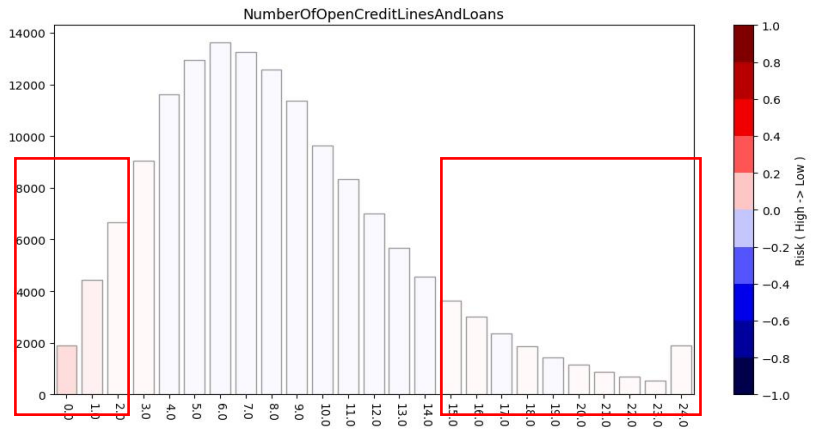
3. RevolvingUtilizationOfUnsecuredLines field has some relatively linear relationship with risk. When RevolvingUtilizationOfUnsecuredLines is low, it is low risk, while high RevolvingUtilizationOfUnsecuredLines means high risk

4. Age field is not significant feature compared with RevolvingUtilizationOfUnsecuredLines field. It seems that young age is relatively high compared with old age customers.

04

Risk Distribution





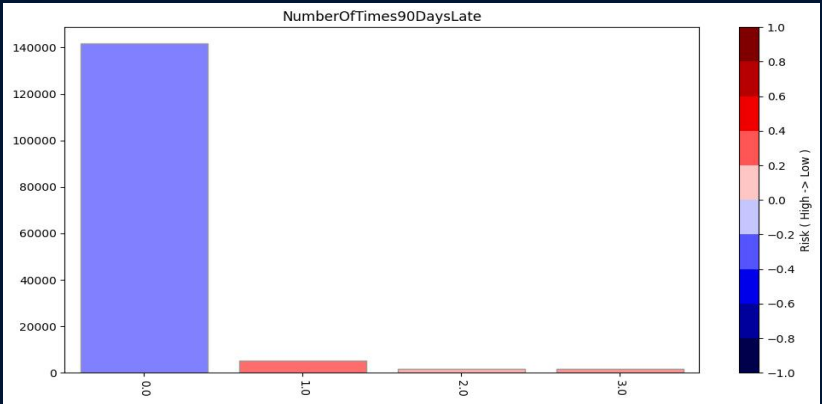
Note:

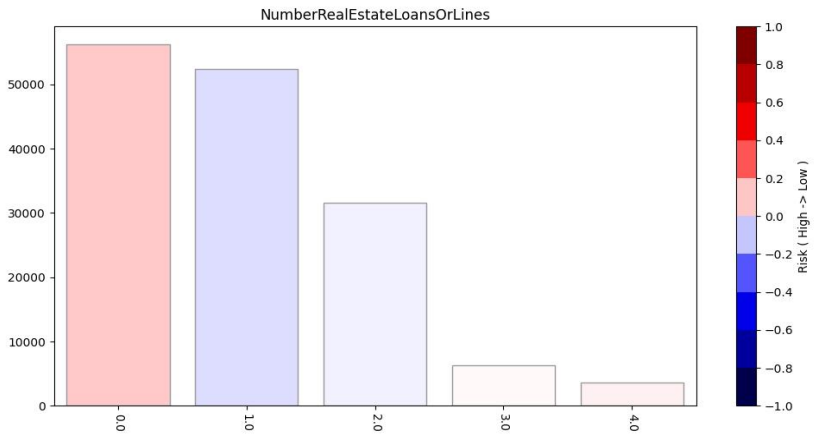
1. NumberOfTime30-59DaysPastDueNotWorse field has low risk at 0 value while relatively high when > 1. It is also interesting that risk decreases when NumberOfTime30-59DaysPastDueNotWorse is over 3.

2. NumberOfOpenCreditLinesAndLoans field is relatively high when the value is less than 3 or larger than 15. It is very interesting.

04

Risk Distribution





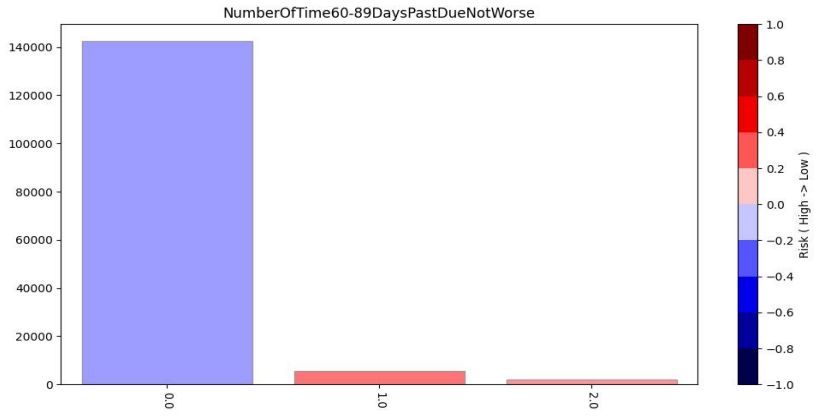
Note:

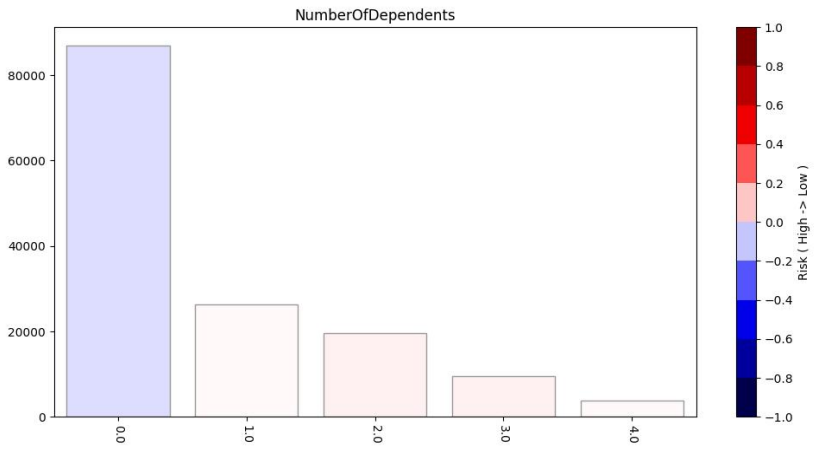
1. NumberOfTimes90DaysLate is at highest risk when the value is 1. Similar to NumberOfTime30-59DaysPastDueNotWorse field.

2. NumberRealEstateLoansOrLines field is high risk at 0. The risk is relatively low at 1 or 2, then increase again.

04

Risk Distribution





Note:

1. NumberOfTime60-89DaysPastDueNotWorse field has low risk at 0 while highest at 1. Similar to other 2 field : NumberOfTimes90DaysLate and NumberOfTime30-59DaysPastDueNotWorse

2. NumberOfDependents filed is relatively low at 0 while relatively high when the number increases. The highest point is at 2 / 3.

05

Xgboost Model - Train

Note:

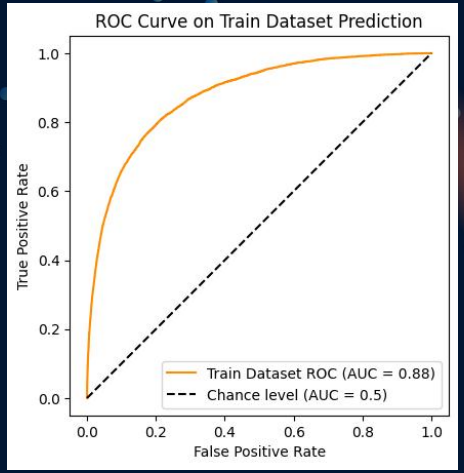
The reason why I selected Xgboost model as the main model here is that Xgboost model has the advantages:
1. As one of the most famous ensemble models, it is relatively has better performance compared with single CART tree model.

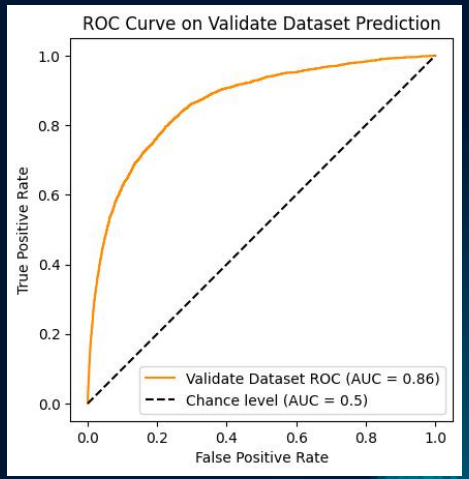
2. Xgboost model can handle the NAN values.

3. Xgboost model can effectively overcome the overfiting compared with other models.

4. Here we also combined with RandomizedSearchCV module that uses Cross-Validation to overcome overfitting.

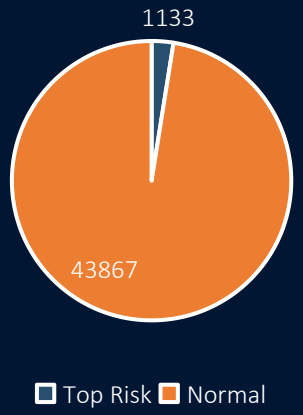
5. The performance of the model is quite good with AUC score at 0.88 on Training Dataset and 0.86 on Validate Dataset (30% of the whole Train Dataset.)





06

High-Risk Validate



	True	False
Positive	631	502
Negative	2287	41580

Note:

We set threshold as 0.5 and select the high-risk customers from **Validate** dataset. This account for 2.5% of the validate dataset.

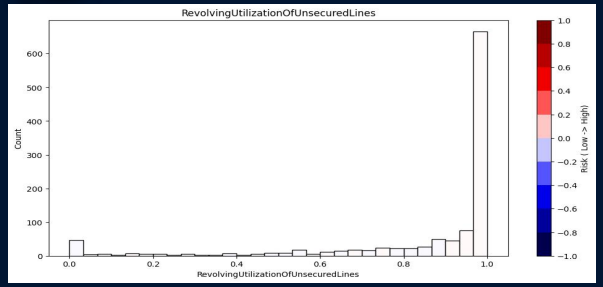
The measurement:

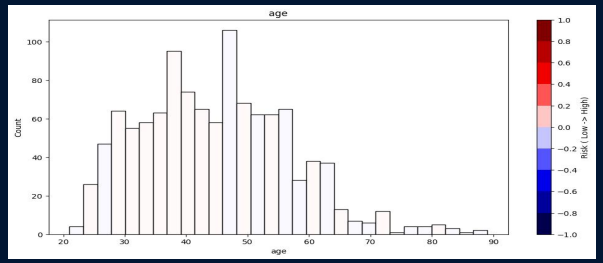
Precison = 631 / (631 + 502) = 55.7%

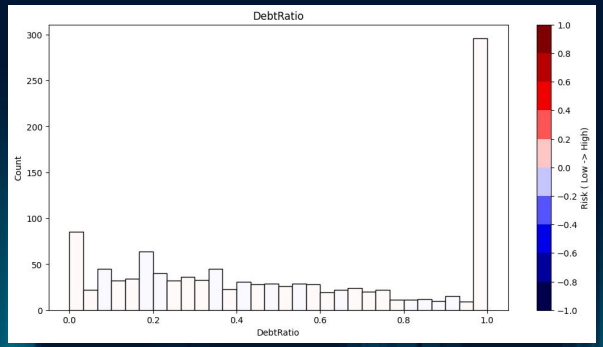
Recall Rate = 631 / (631 + 2287) = 21.6%

06

High-Risk Validate







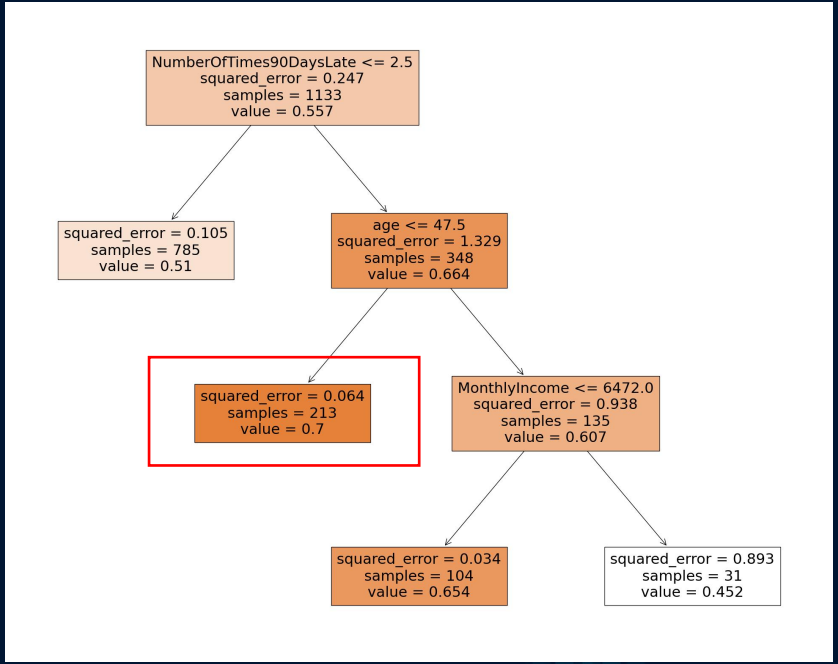
Note:

Based on the risk distribution on different features, we can see that the risk distribution is quite uniformly distributed.

This partly prove that our model is working well.

06

High-Risk Validate



Note:

To better refine the result, here we use CART tree to split the current customer pool based on the features.

It shows that NumberOfTimes90DaysLate field is the first factor that spit the customers, then age and MonthlyIncome.

The Highest risk score here is the branch:

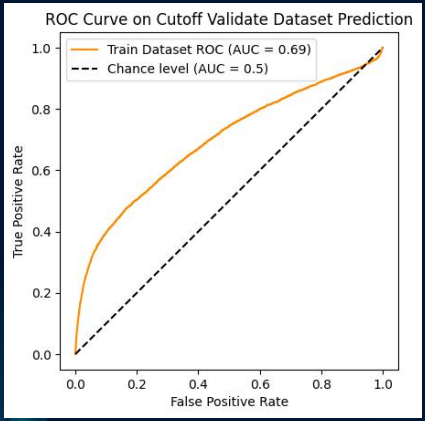
(NumberOfTimes90DaysLate > 2.5) & (age <= 47.50)

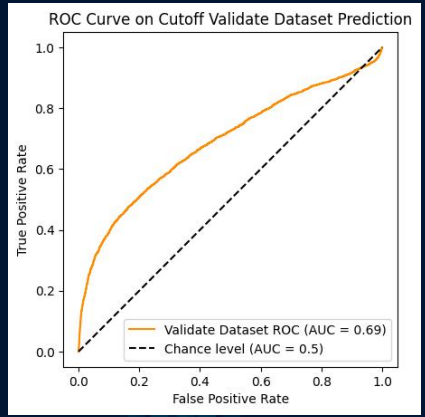
Precision : 70%

Logistic Regression Model

4.2 PCA preprocess to remove non-linearity

```
pca = PCA(n_components=len(features))
lrX_train_pca = pca.fit_transform(lrX_train)
lrX_validate_pca = pca.transform(lrX_validate)
```





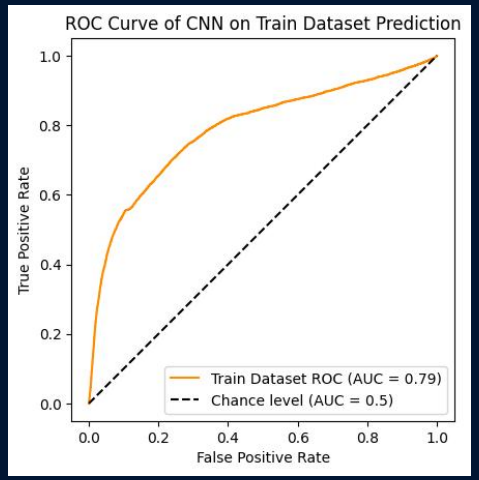
Note:

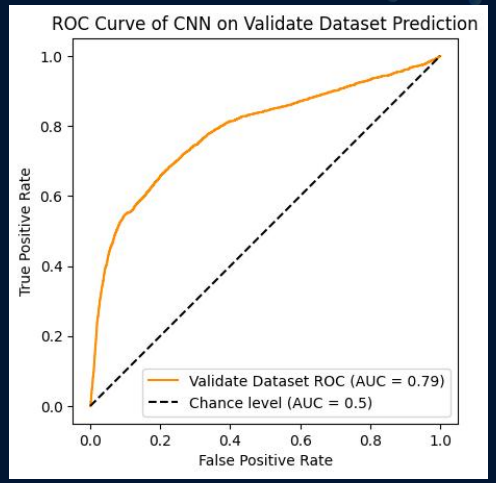
Since there are some collinearity among the features (as we showed in Part 3), we applied PCA to the data first.

The AUC score is 0.69 for Train Dataset and Validate Dataset.

08

CNN Model





Note:

Here we applied CNN model as upgrade compared with Logistic Regression model.

We can treat CNN as emsemble model of multiple Logistic Regression models.

The performance here is 0.79 on both Train and Validate dataset.

09

Conclusion

Note:

In this case, we used Xgboost, Logistic Regression and CNN model to predict the risk score of the customers.

Based on the performance, Xgboost > CNN > Logistic Regression.

As for the features, RevolvingUtilizationOfUnsecuredLines , NumberOfTimes90DaysLate , NumberOfTime60-89DaysPastDueNotWorse, NumberOfTime30-59DaysPastDueNotWorse are the most important features.

If we set a threshold as 0.5 and we can make the result better with some rules extracted from tree models

Thank you

Email: zhang.ai.japan@gmail.com