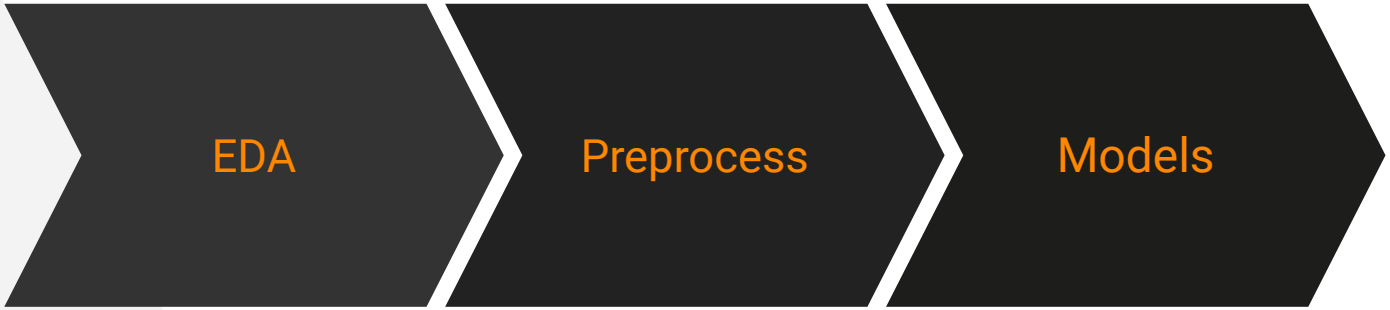
Comparative Model Analysis

Jiacheng Yao

Process



1. Exploratory Data Analysis

Exploratory Data Analysis

Application

Dimension: 276686, 122

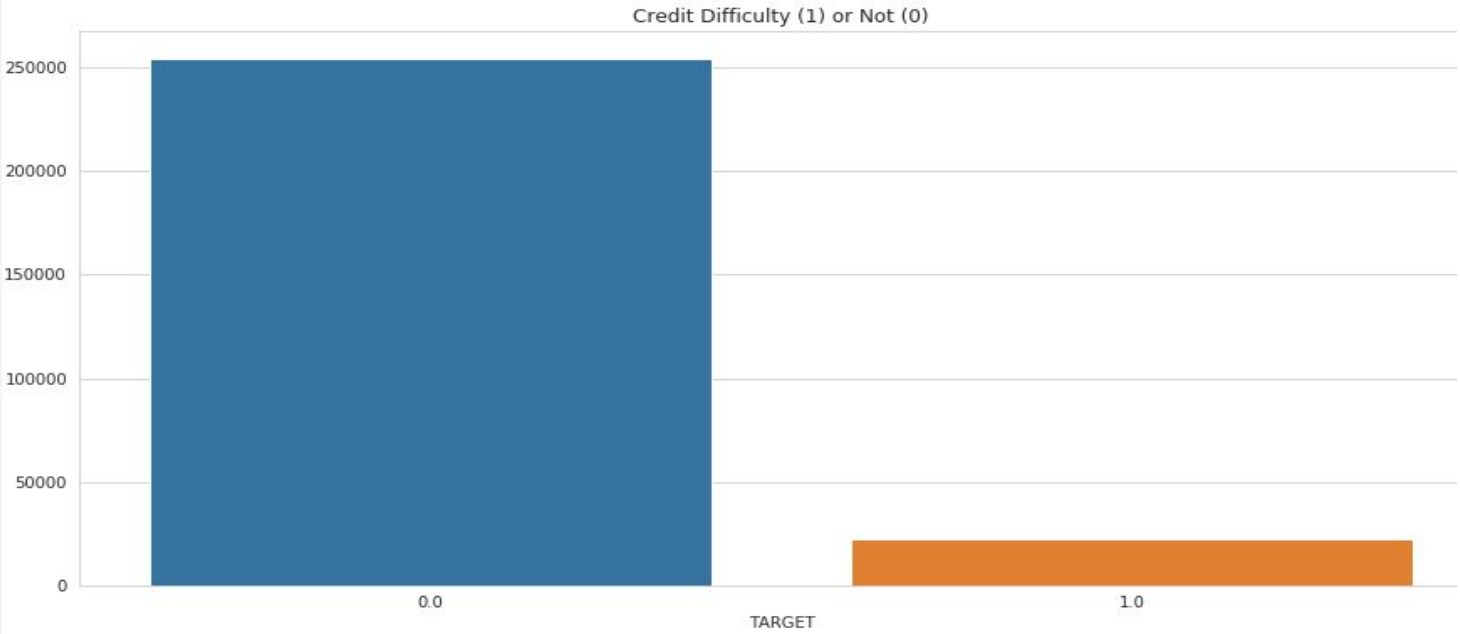
One row = One loan application

Bureau

Dimension: 1716428, 17

One row = One previous Loan

Exploratory Data Analysis



2. Preprocess

Preprocess – Merge Bureau Data

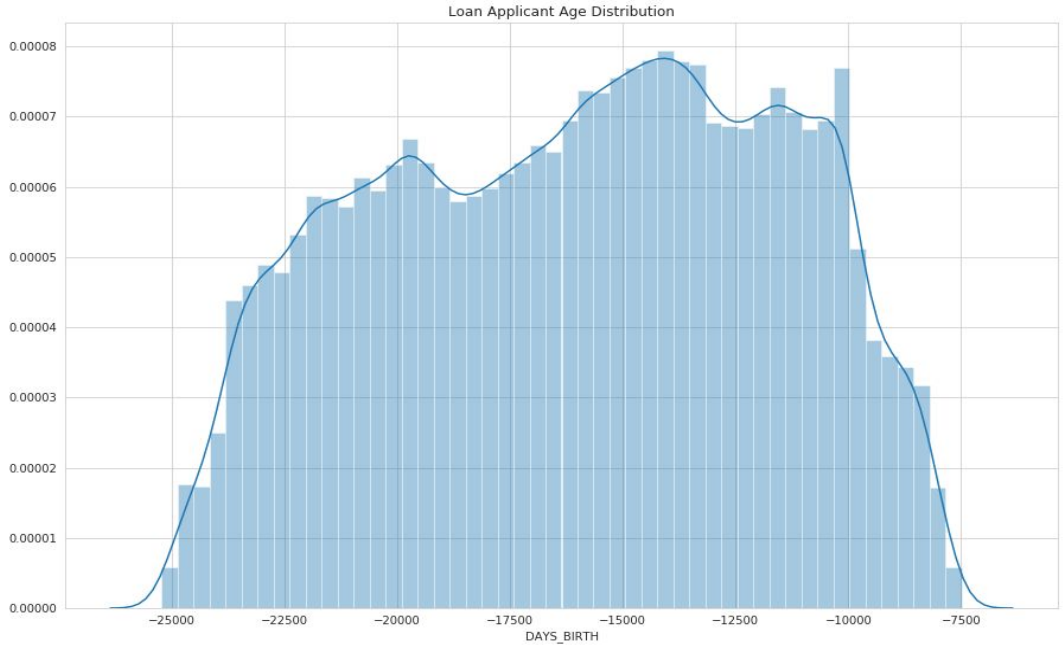
Categorical

Aggregate to Most Common per SK_ID_CURR

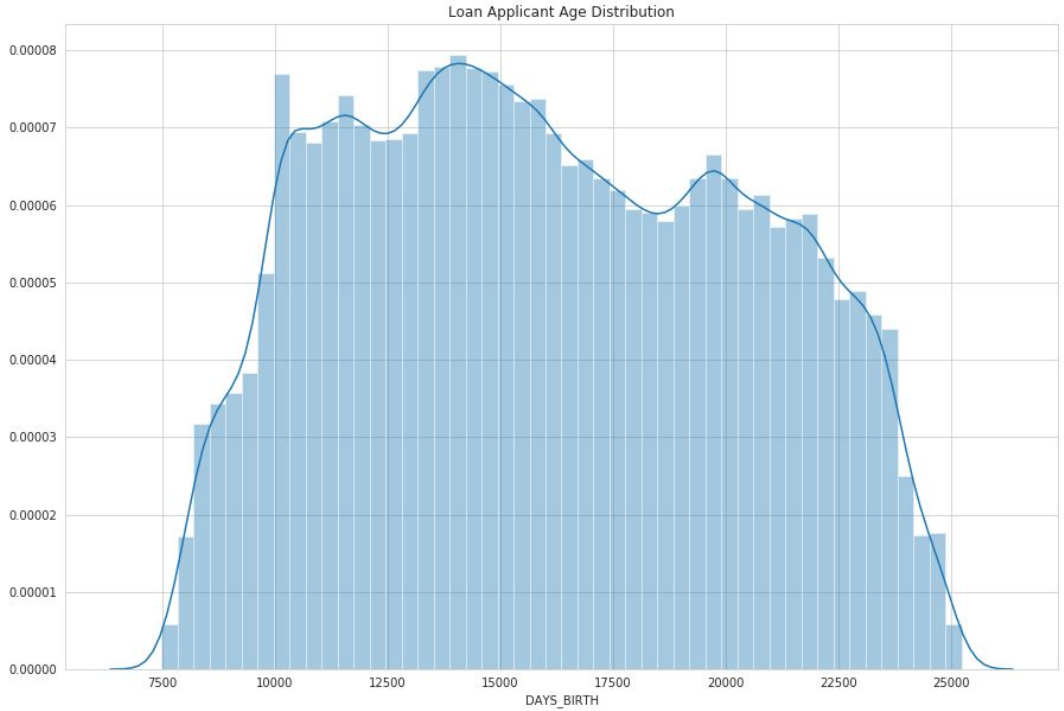
Numeric

Aggregate to Median per SK_ID_CURR

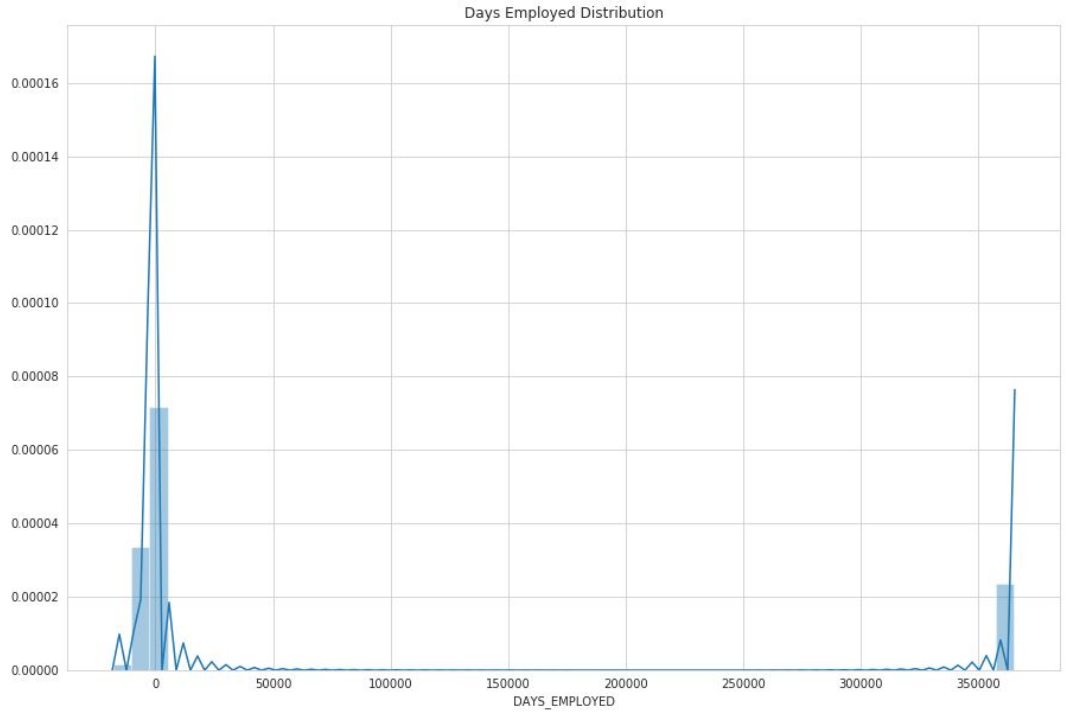
Preprocess - Loan Applicant Age



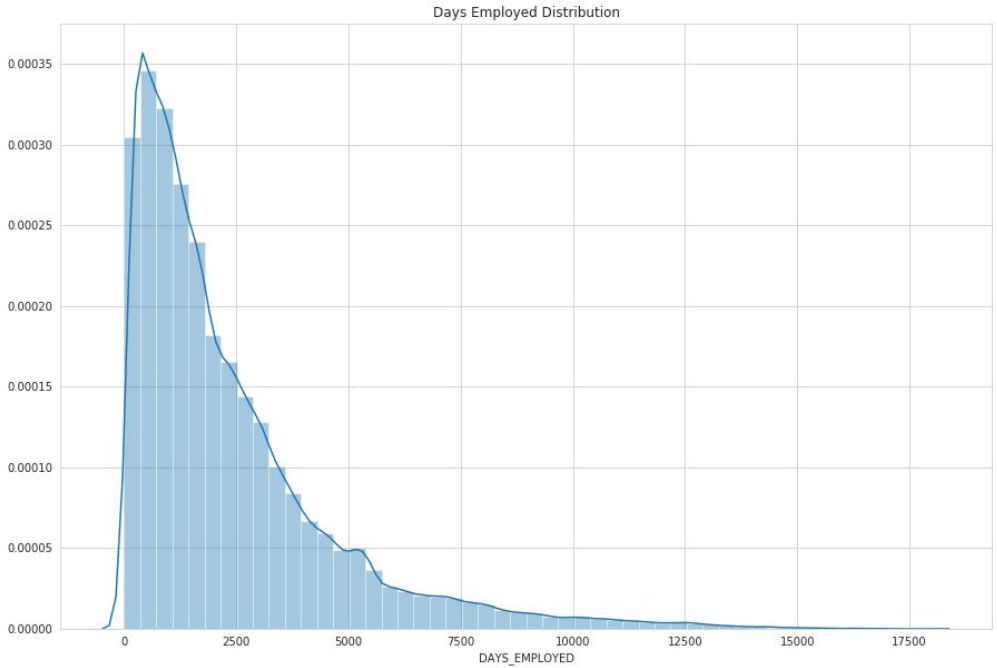
Preprocess - Loan Applicant Age (clean)



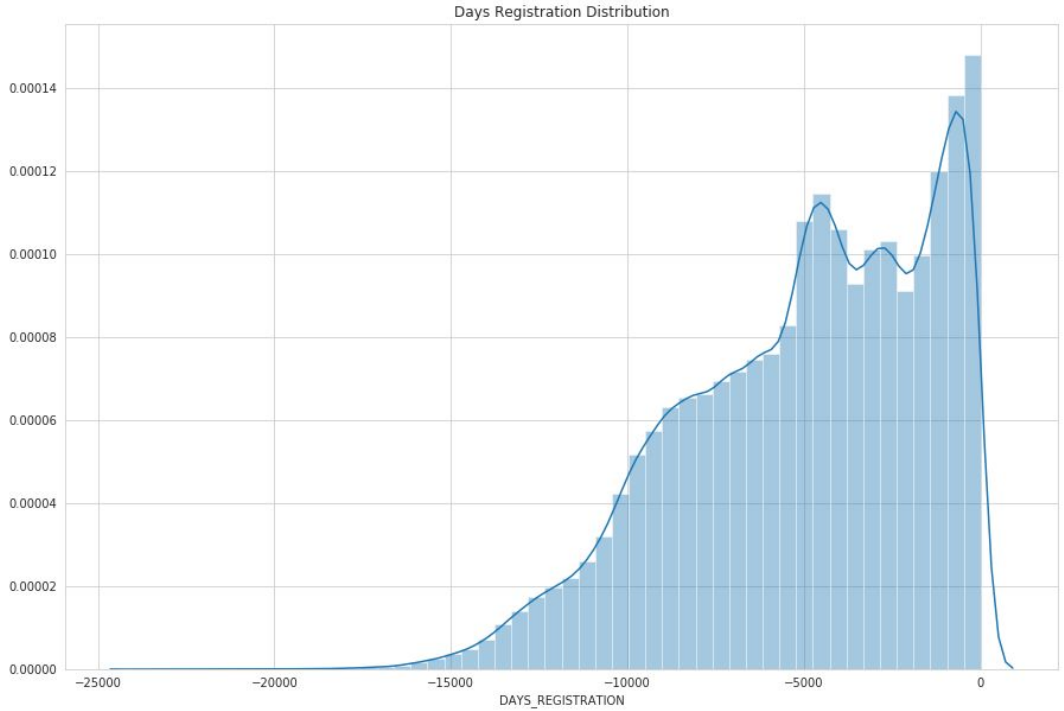
Preprocess - Days Employed



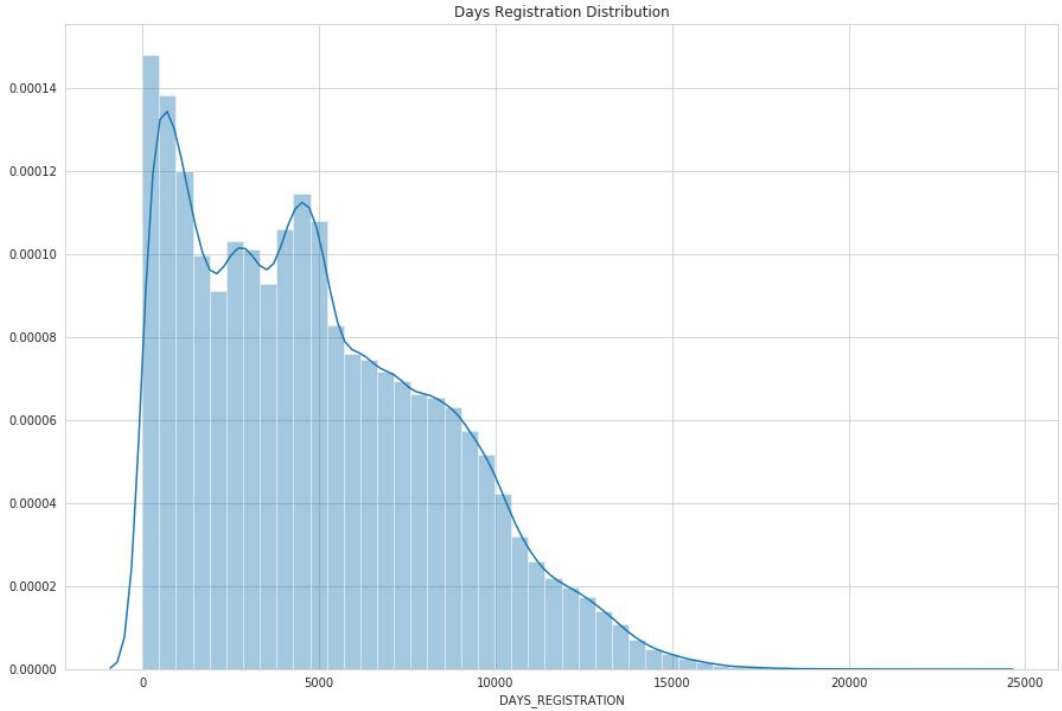
Preprocess - Days Employed (clean)



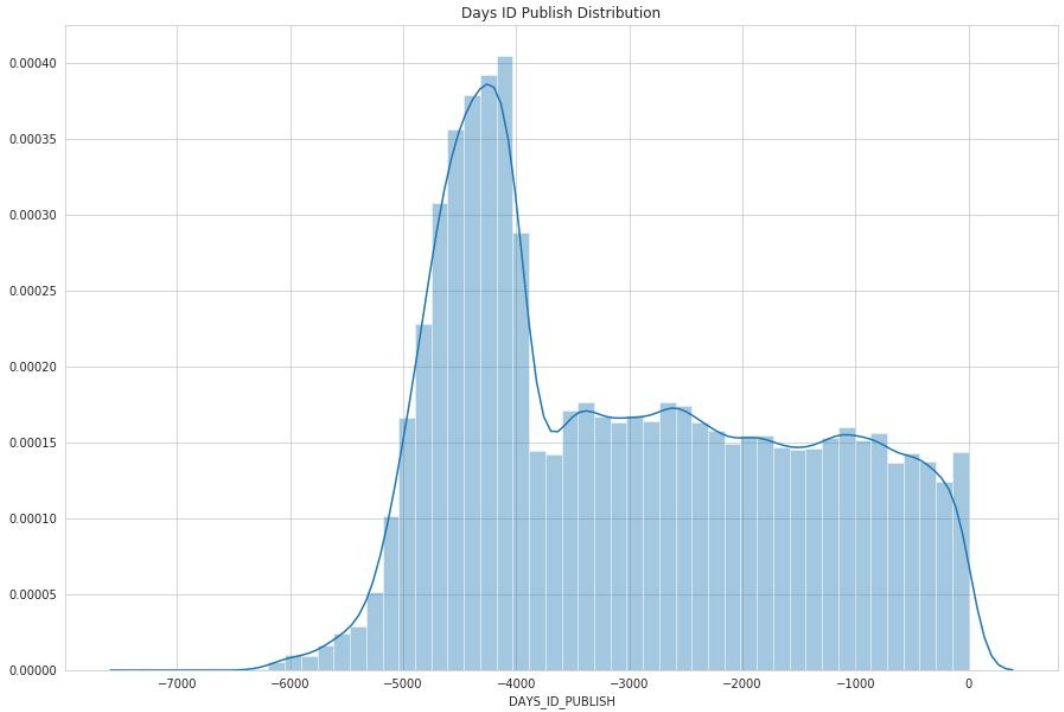
Preprocess - Days Registration



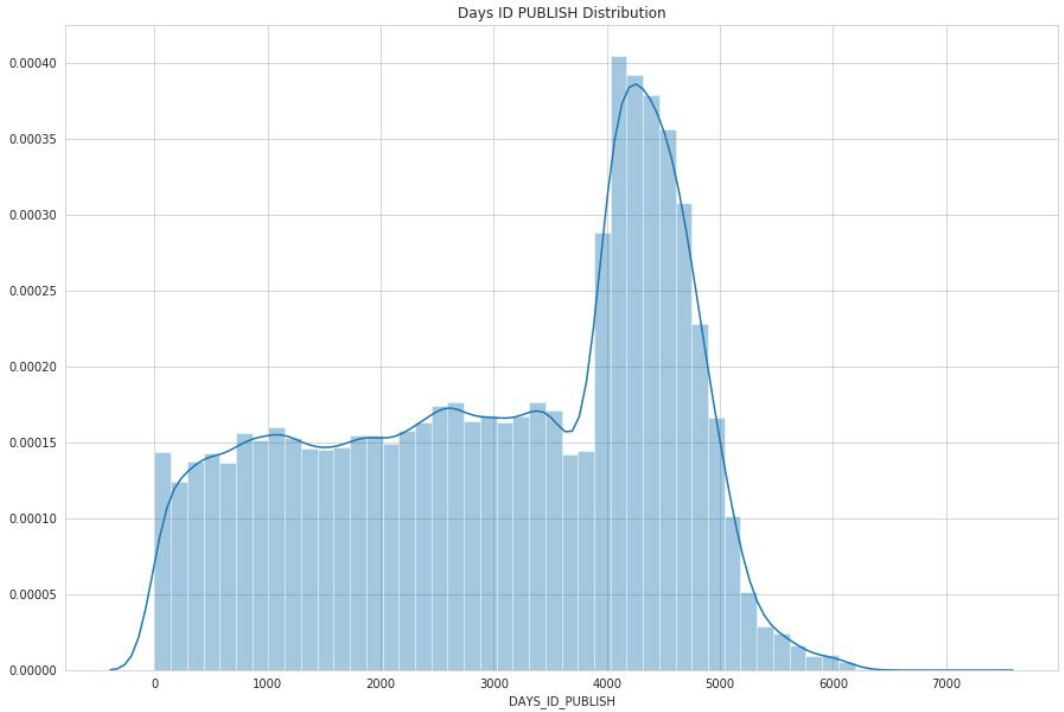
Preprocess - Days Registration (clean)



Preprocess – Days ID Public



Preprocess - Days ID Public (clean)



Preprocess - Last Steps

Categorical

1. Null -> “Missing”
2. One Hot Encoding

Numeric

1. Drop features with high correlation
2. Impute with MICE
3. Standardize

3. Model Comparison

Model Comparison

	Fast and Frugal	Random Forest	XgBoost
BACC (10Fold AVG)	**0.6805**	**0.6808**	**0.8129**
Time	**12’20”**	**14’26”**	**71’18”**

4. Discussion

Discussion

Simplicity and Interpretability

FFT: more time efficient, interpretable, human readable and usable, potentially more robust against change of circumstances (regime changes).

Others: more complicated, difficult to interpret, prone to overfit, less robust, more time-consuming

When to use FFT

1. If used by human without aid,
2. If reducing time and cost is crucial,
3. If understanding and transparency is important,
4. If performance similar to more complicated models,
5. If robustness is priority.

Discussion

Further Steps:

1. Closer inspection of individual features,
2. Integration of alternative data, e.g. credit card data, loan installment payment data, etc,
3. Acquire cost matrix for each prediction scenario (false negative, true negative, false positive, true positive) from experts and optimize models on it to tackle imbalancedness of the dataset.
4. etc.

Behavioral Interventions with ML

Jiacheng Yao

Approaches - Unsupervised

Unsupervised

1. Treat as clustering problem, group debtors into clusters and adopt different interaction strategies for different clusters.
2. To find the optimal strategy for each individual cluster, conduct A/B test on several strategies and choose the one with the best KPI (debt repayment percentage, click through rate, customer retention rate, etc.)

Approaches - Supervised

Classification

1. Treat as classification problem, engineer a target signal to indicate if debtors will continue to repay, or will have difficulty with payments.
2. For debtors with and without difficulty paying back debts, adopt different interaction strategies for the two classes.
3. To find the optimal strategy for each individual class, conduct A/B test on several strategies and choose the one with the best KPI (debt repayment percentage, click through rate, customer retention rate, etc.)

Regression

1. Treat as regression problem, engineer a target signal to forecast future lifetime value(LV) of debts.
2. Divide debtors into different tranches based on future LV, adopt different interaction strategies for different tranches.
3. To find the optimal strategy for each individual tranch, conduct A/B test on several strategies and choose the one with the best KPI (debt repayment percentage, click through rate, customer retention rate, etc.)

Recommendations

1. Monitor the KPIs vigilantly with incoming data to validate the performance of the models and do optimization based on results.
2. If possible, conduct surveys to better gauge satisfaction levels of different groups of debtors and how to improve interaction methods.
3. Integrate findings from experts in the field of behavioral science, especially in the use case of debt collection.
4. etc.

THANKS!

Any questions?

You can find me at jc07.yao@gmail.com