Credit Default Risk Analysis

Lina Gao

Problem Statement

- *How do we use modern machine learning algorithms to identify potential credit defaulter based on their historical transaction data and socioeconomic status?*

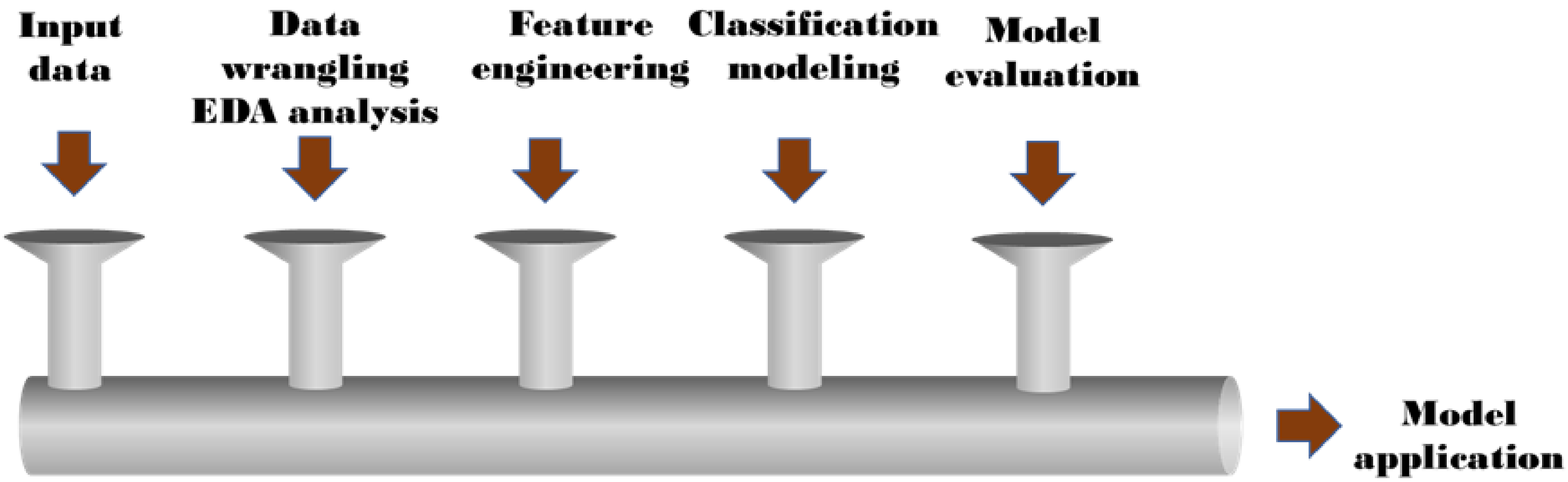
Source of Data

- It includes 30000 observations with 25 features collected in Taiwan from April 2005 to Sep 2005.
- Features: Age, Sex, Marriage, Education, credit limit, payment, bill and payment status in each of the six months.

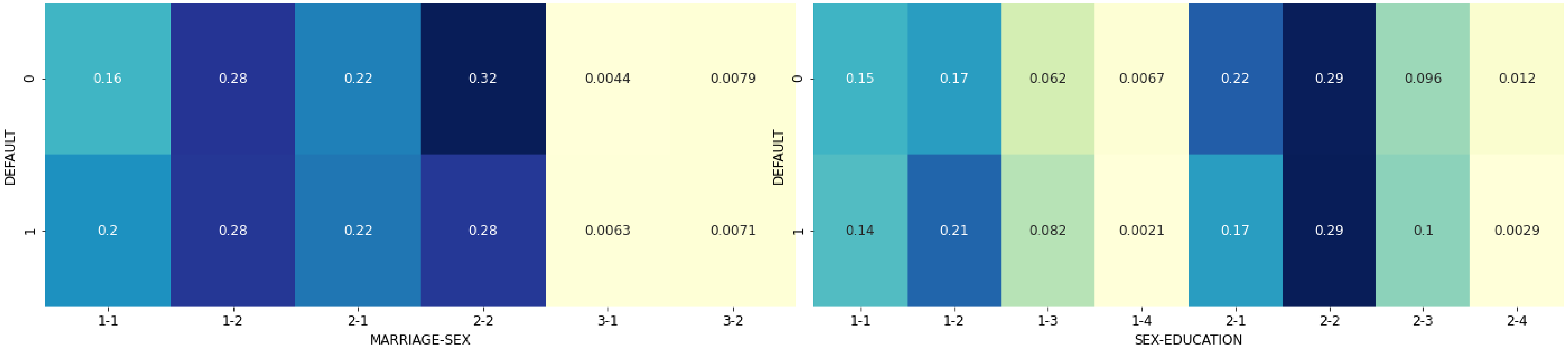
Data structure

Variables	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
MARRIAGE	1=married, 2=single, 3=other
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
DEFAULT	1=default, 0=non-default. * N(non-default):N(default)=3.52:1
PAY_0, PAY_2, …, PAY_6	Repayment status in September, August, July, June, May, April (-2=no consumption, -1=pay duly, 0=the use of revolving credit, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
BILL_AMT1, BILL_AMT2, …, BILL_AMT6	Amount of bill statement in September, August, July, June, May, April, 2005 (NT dollar)
PAY_AMT1, PAY_AMT2, …, PAY_AMT6	Amount of previous payment in September, August, July, June, May, April 2005 (NT dollar)

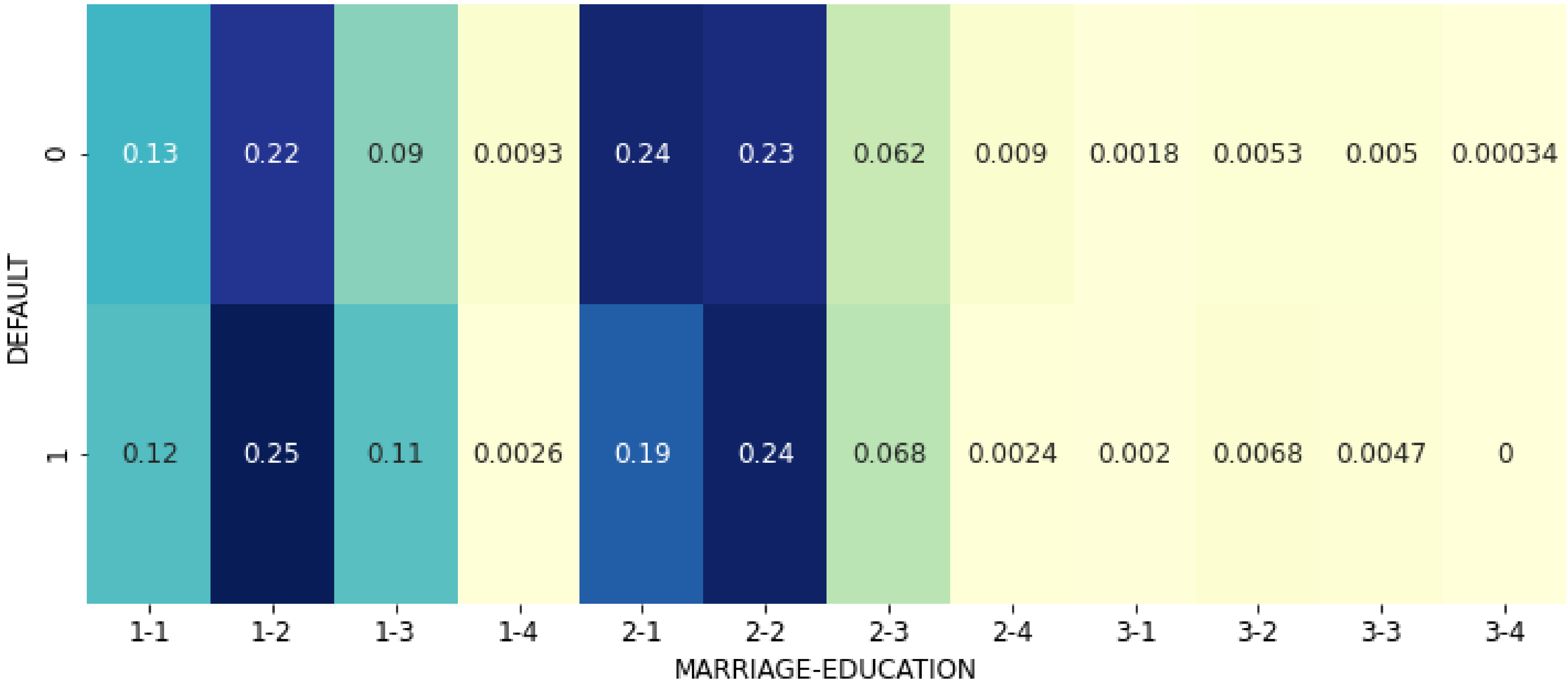
Analytical Workflow



Data Exploration: client segmentation



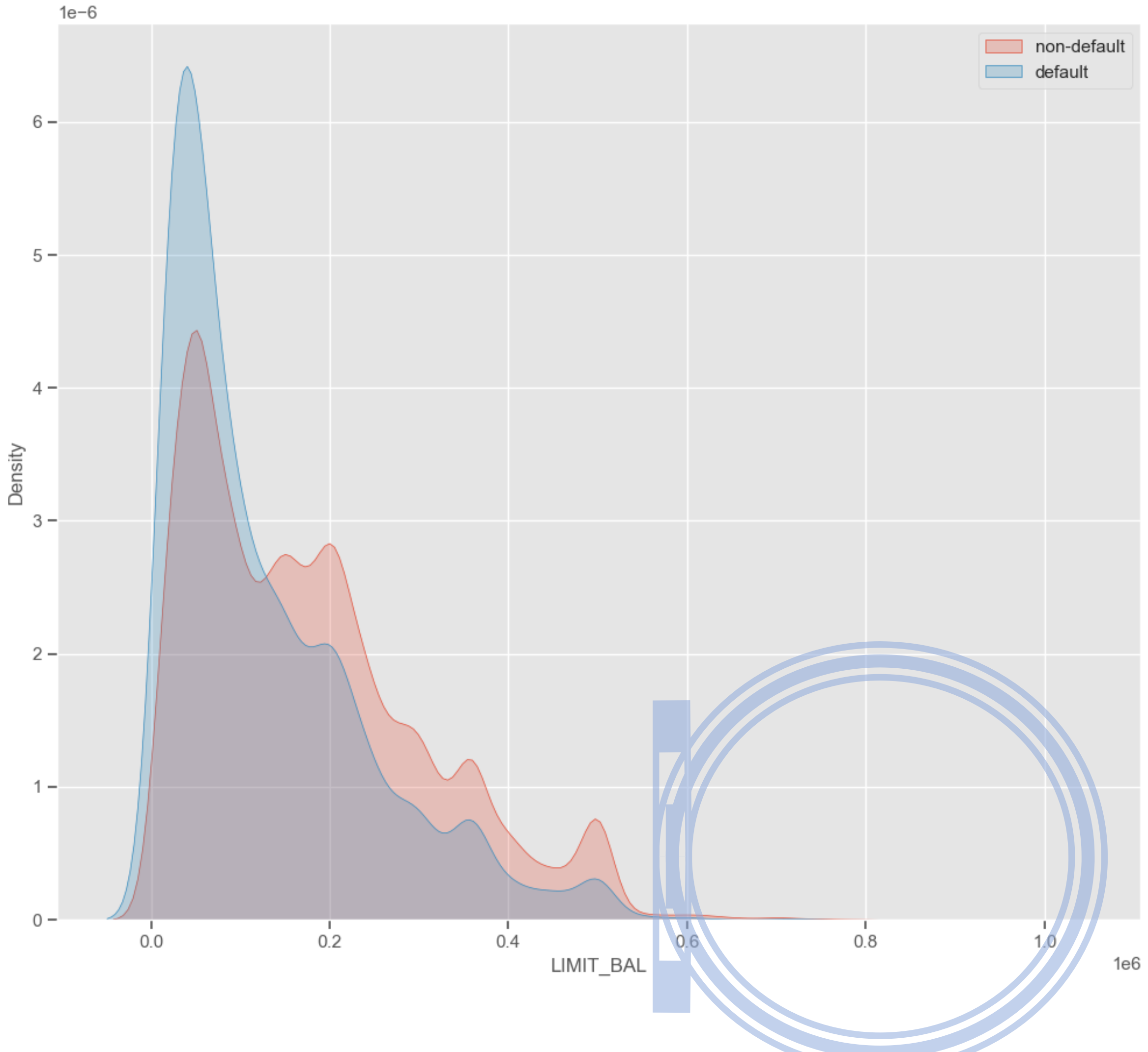
Data Exploration

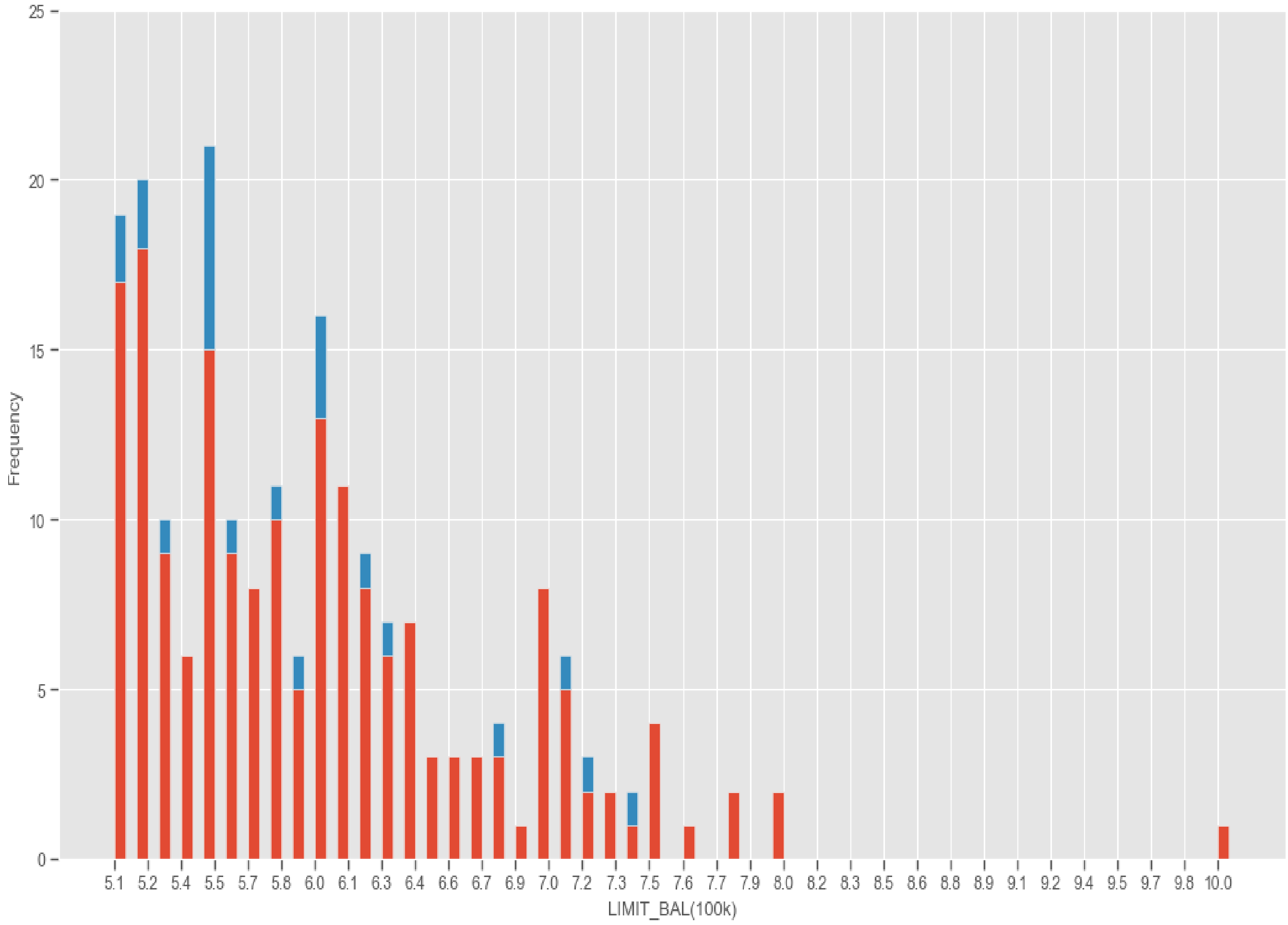


Statistics test to explore the relationship between the features

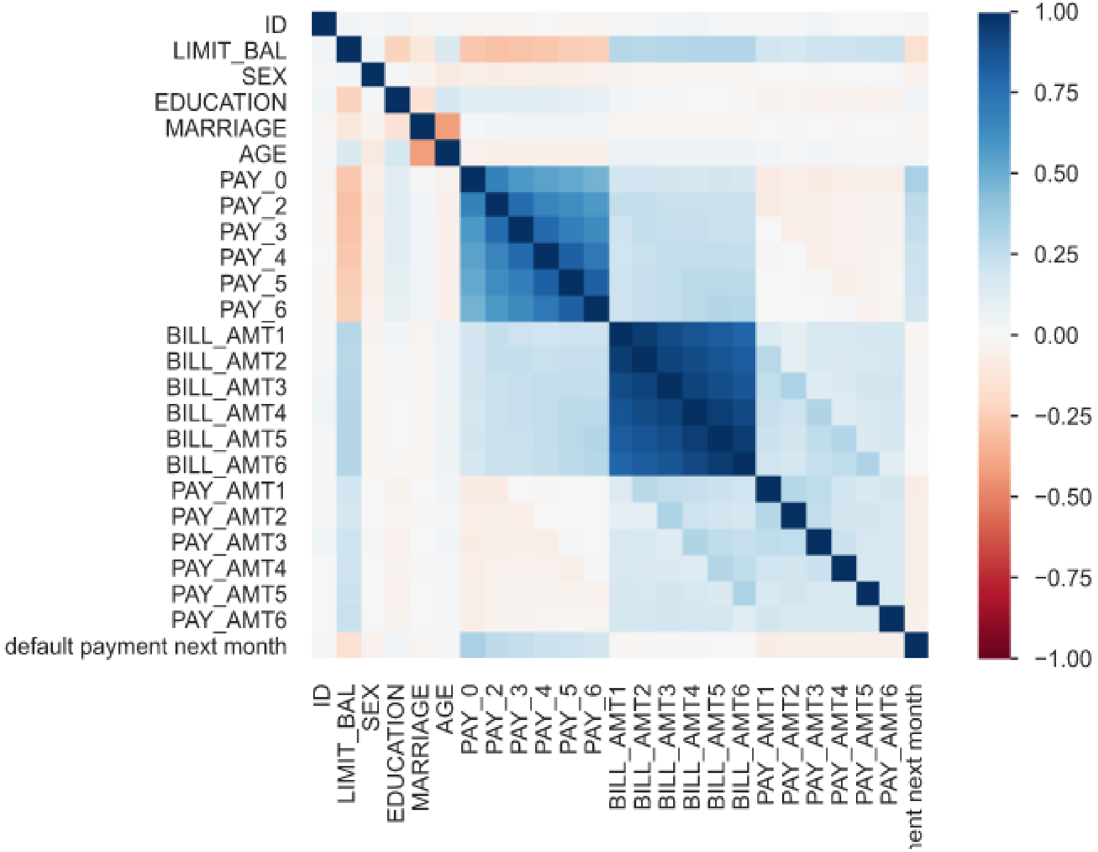
features	Statistic tests	Results
DEFAULT vs SEX	Chi-square test	P<0.05
DEFAULT vs MARRIAGE		
DEFAULT vs EDUCATION		
LIMIT_BAL vs SEX	Mann-Whitney U test and t test	P<0.05
LIMIT_BAL vs MARRIAGE		
LIMIT_BAL vs EDUCATION	Kruskai-Wallis test and ANOVA	P<0.05
LIMIT_BAL vs DEFAULT	Mann-Whitney U test and t test	P<0.05

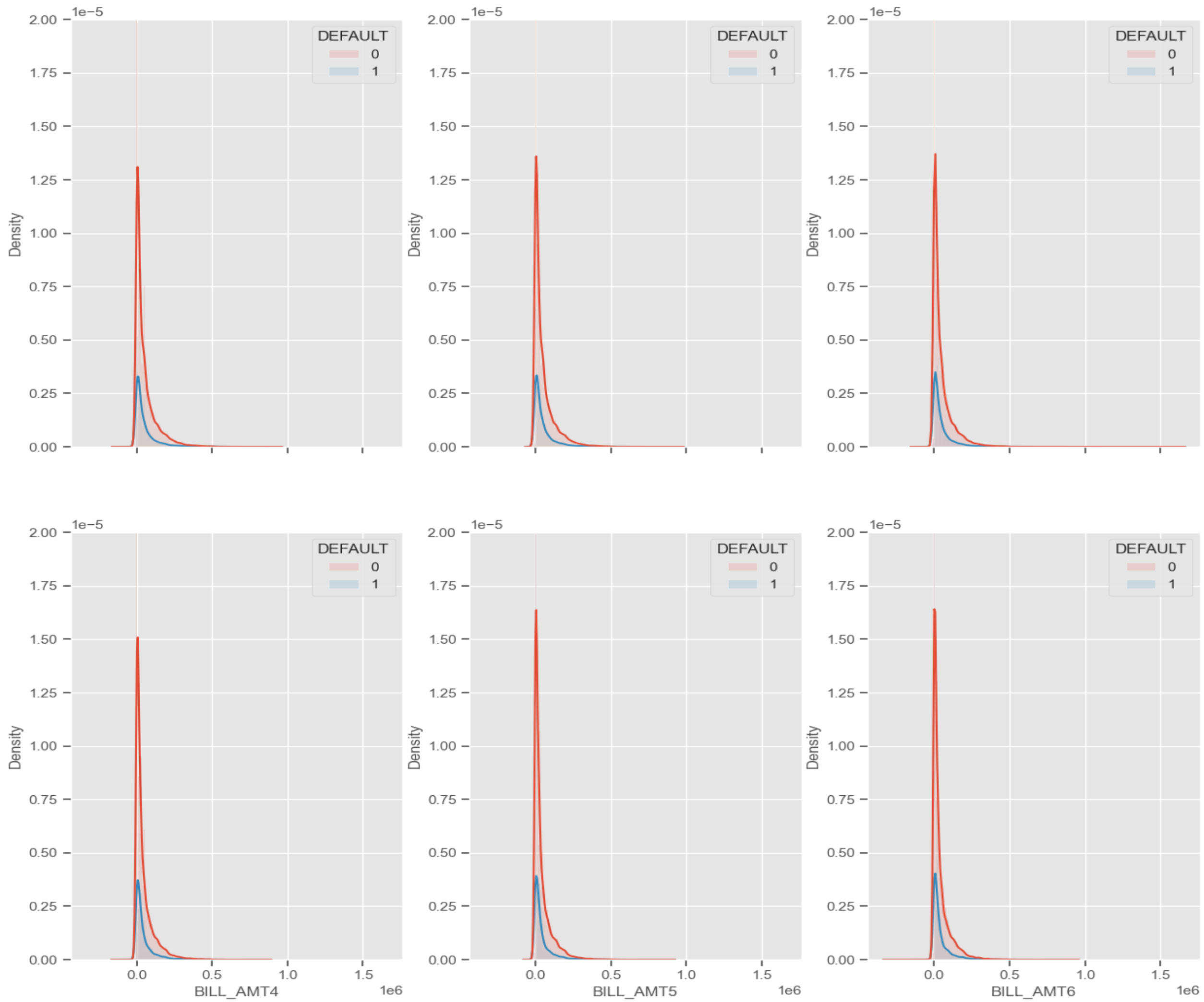
Data Exploration





Data Exploration

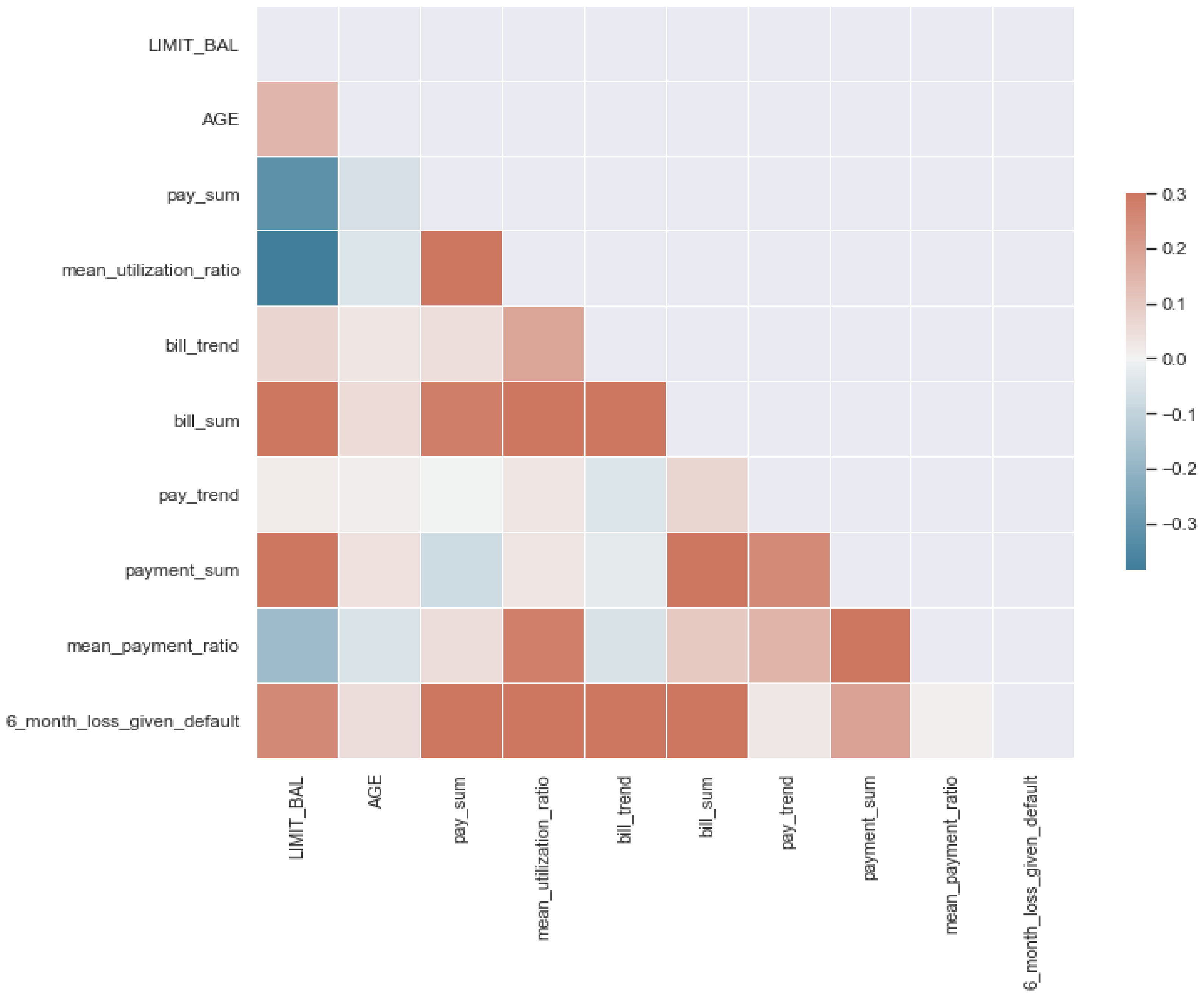




Feature Engineering

1. $pay\_sum = \sum_{i=1}^{6} PAY\_i$
2. $mean\_utilization\_ratio = \sum_{i=1}^{6} BILL_{AMTi} /(6 \times LIMIT\_BAL)$
3. $6\_month\_loss\_given\_default = \sum_{i=1}^{6} BILL\_AMTi - \sum_{i=1}^{6} PAY\_AMTi$
4. $mean\_payment\_ratio = \sum_{i=1}^{6} PAY\_AMTi \ /(6 \times LIMIT\_BAL)$
5. $bill\_trend = (\sum_{i=1}^{3} BILL\_AMTi - \sum_{i=4}^{6} BILL\_AMTi \ )/(3 \times LIMIT\_BAL)$
6. $pay\_trend = (\sum_{i=1}^{3} PAY\_AMTi - \sum_{i=4}^{6} PAY\_AMTi \ )/(3 \times LIMIT\_BAL)$
7. $bill\_sum = \sum_{i=1}^{6} BILL\_AMTi$
8. $payment\_sum = \sum_{i=1}^{6} PAY\_AMTi$

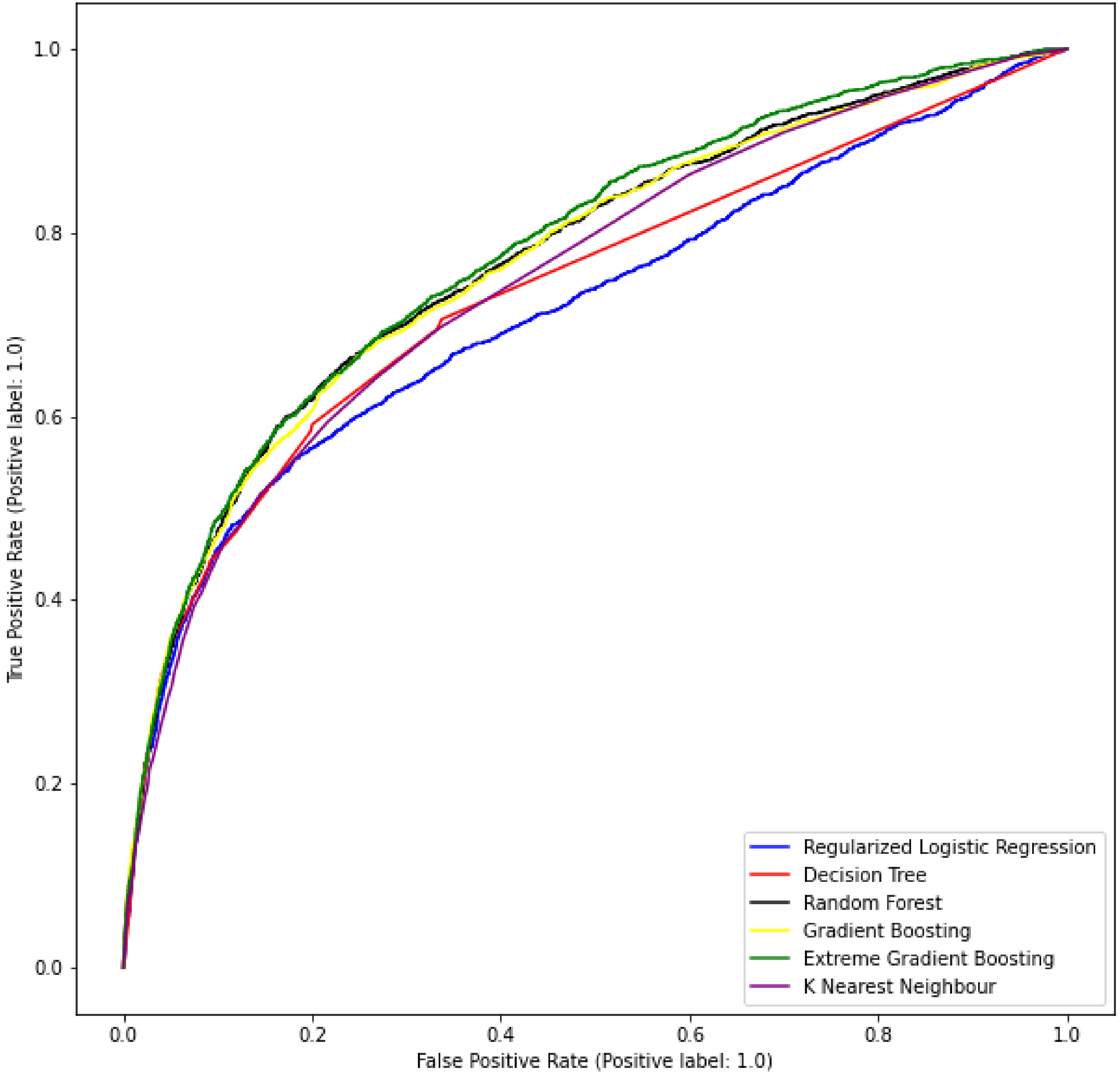
Feature Engineering



Modeling using the original data set

Algorithms	Hyperparameters	best estimate	accuracy	precision	recall	f1 score	AUC	runtime	top 3 most important features
regularized logistic regression	C	50	0.809	0.68	0.24	0.357	0.716	7.34	PAY_1, BILL_AMT1, PAY_AMT1
decision tree	criterion, max_depth	entropy, 4	0.817	0.663	0.352	0.459	0.716	19.99	PAY_1, PAY_2, PAY_AMT3
random forest	n_estimators, max_depth	200, 6	0.816	0.662	0.342	0.451	0.771	109.33	PAY_1, PAY_2, PAY_3
gradient boosting	learning_rate, n_estimator, max_depth	50, 2, 0.1	0.819	0.673	0.351	0.462	0.768	1814.9	PAY_1, PAY_2, PAY_5
Extreme gradient boosting	subsamples, n_estimators, max_depths, learning rate, gamma, reg_alpha	0.5, 200, 5, 0.05, 0.1, 0	0.818	0.663	0.361	0.468	0.779	464	PAY_1, PAY_1, PAY_3
KNN	n_neighbors, p	50, 2	0.806	0.652	0.265	0.377	0.749	85	PAY_1, PAY_2, PAY_3

ROC curves comparison



Modeling using engineered features

Algorithms	Hyperparameters	best estimate	accuracy	precision	recall	f1 score	AUC	top 3 most important features
regularized logistic regression	C	0.5	0.797	0.672	0.164	0.264	0.685	pay_sum, payment_sum, LIMIT_BAL
decision tree	criterion, max_depth	entropy, 4	0.804	0.638	0.262	0.371	0.748	pay_sum, payment_sum, bill_sum
random forest	n_estimators, max_depth	200, 9	0.804	0.625	0.28	0.387	0.769	pay_sum, payment_sum, bill_trend
gradient boosting	learning_rate, n_estimator, max_depth	50, 2, 0.1	0.805	0.626	0.293	0.399	0.764	pay_sum, payment_sum, bill_sum
Extreme gradient boosting	subsamples, reg_alpha, n_estimators, max_depths, learning rate, gamma	0.5, 0.0, 200, 5, 0.05,0.1	0.802	0.605	0.305	0.405	0.768	pay_sum, payment_sum, bill_sum
KNN	n_neighbors, p	2, 50	0.799	0.655	0.195	0.301	0.717	pay_sum, payment_sum

Conclusion

- Regardless of the socioeconomic status of the clients, payment status is the most critical feature for credit default prediction. It is described as PAY_1 to PAY_6 in the original data or pay_sum in the reduced data set.
- Features related to a socioeconomic status like age, marriage, and education significantly affect the default in credit risk assessment.
- XGB is the most attractive algorithm for predicting credit default risk compared with RLR, DT, GB, RF and KNN using both original and the reduced data set.
- df_sum using the engineered features gave us comparable results to the original data set.

Limitations

- Data bias: Because this data set is from Taiwan instead of the US, it has limited application reference for consumer credit prediction in the US.
- All these analysis is only applied to the existing clients of the credit card company, not for the prospective ones.
- Features: We didn't have credit bureau data in this project. We also applied historical data, not the information recently.

Future work

Explore

Explore the application of deep learning models in this data set and compare it with the ensemble algorithms.

Use

Use grid search instead of random search to find the optimized hyperparameters and compare the results.

Use

Use resampling technique to balance the data before model development and compare the results.

Develop

Develop a modeling pipeline to extract information more efficiently, providing an automated and faster solution for making credit decisions on time.