Credit Default Risk Analysis

Lina Gao

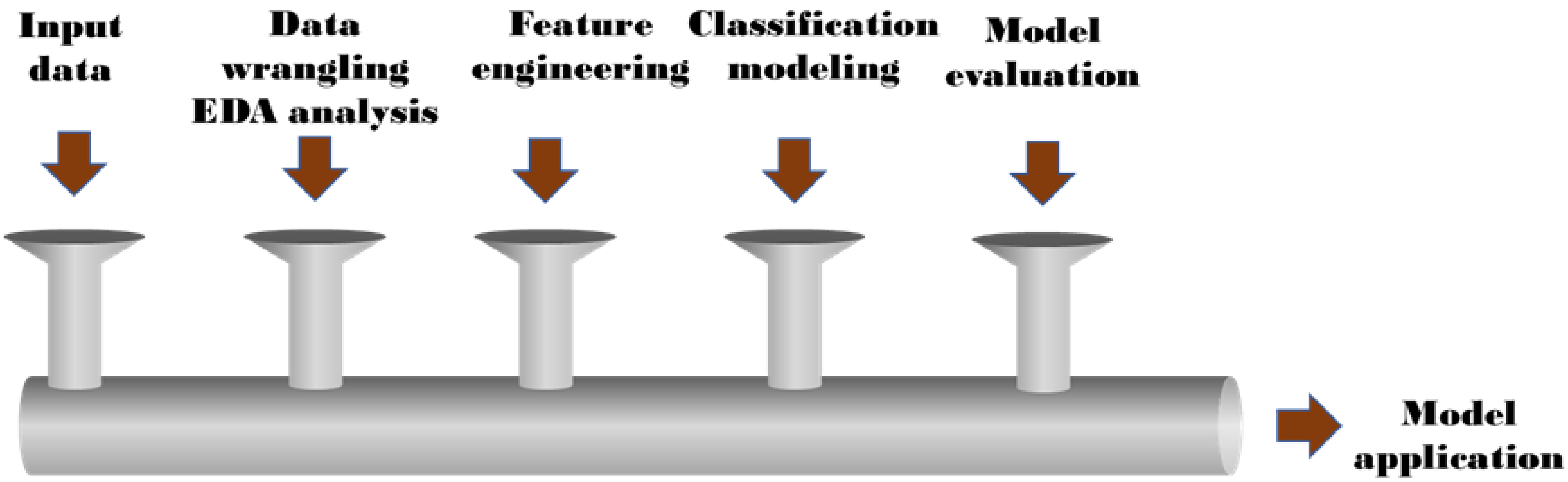
Problem Statement

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

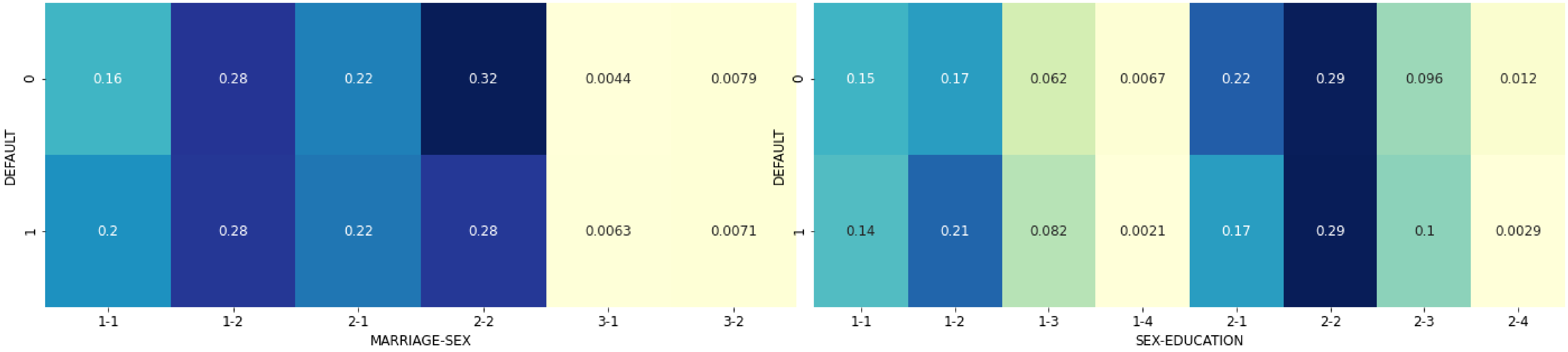
Source of Data

- 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.

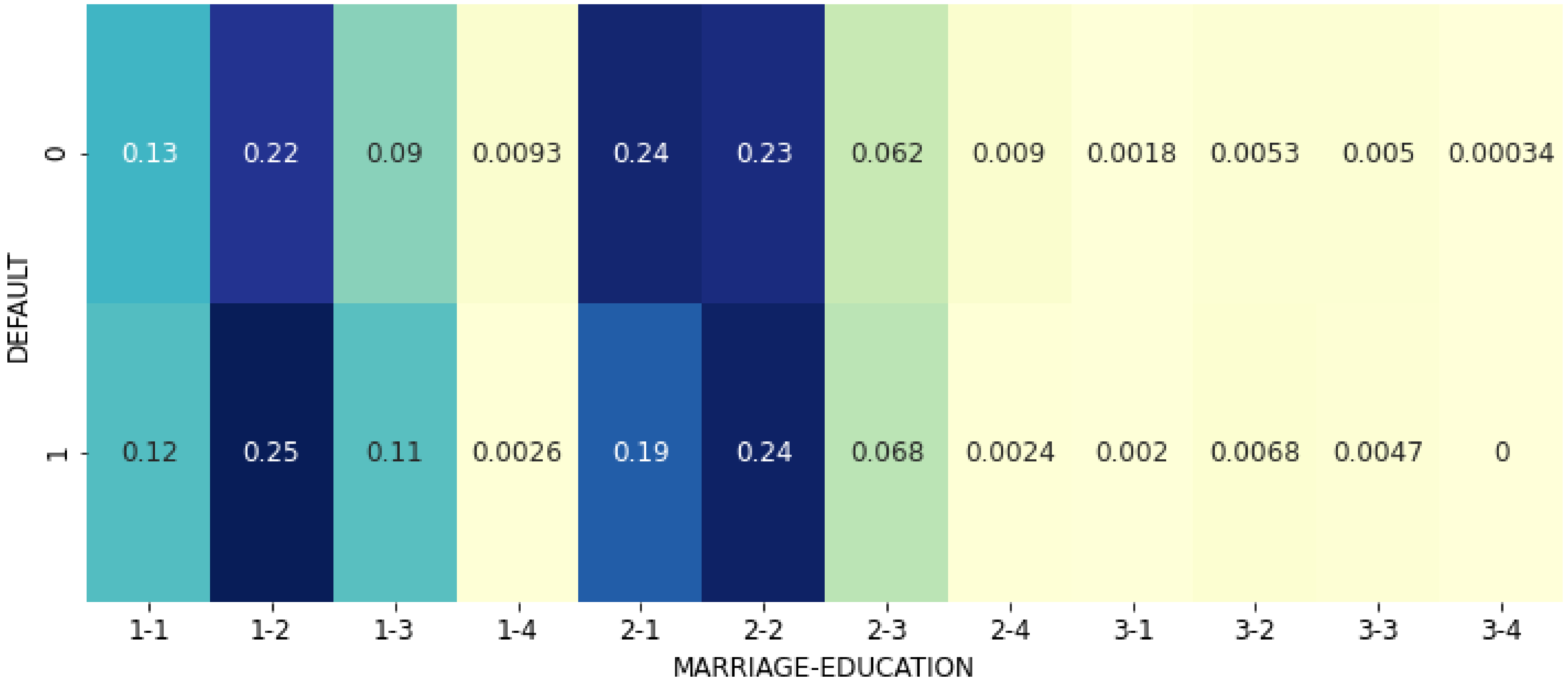
Analytical Workflow



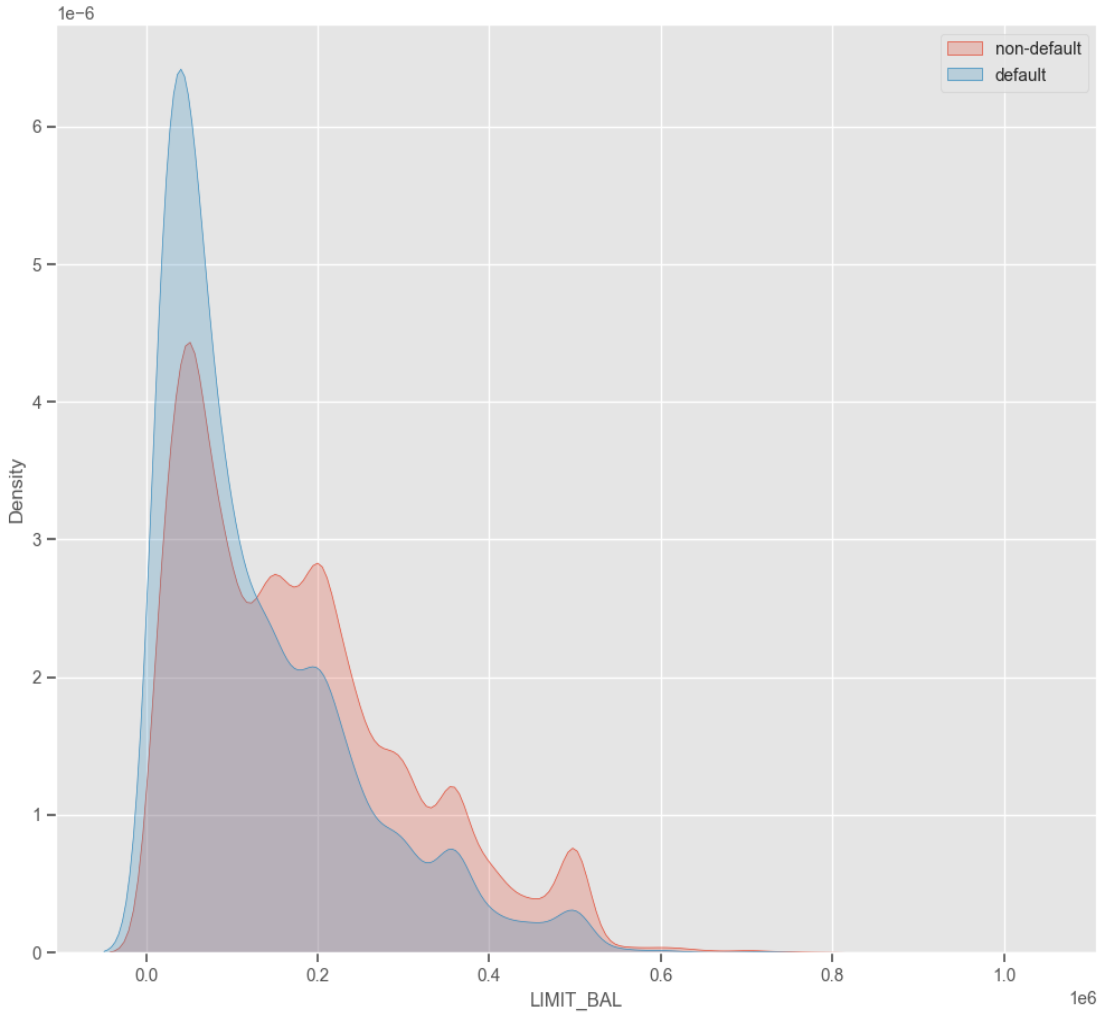
Data Exploration: client segmentation

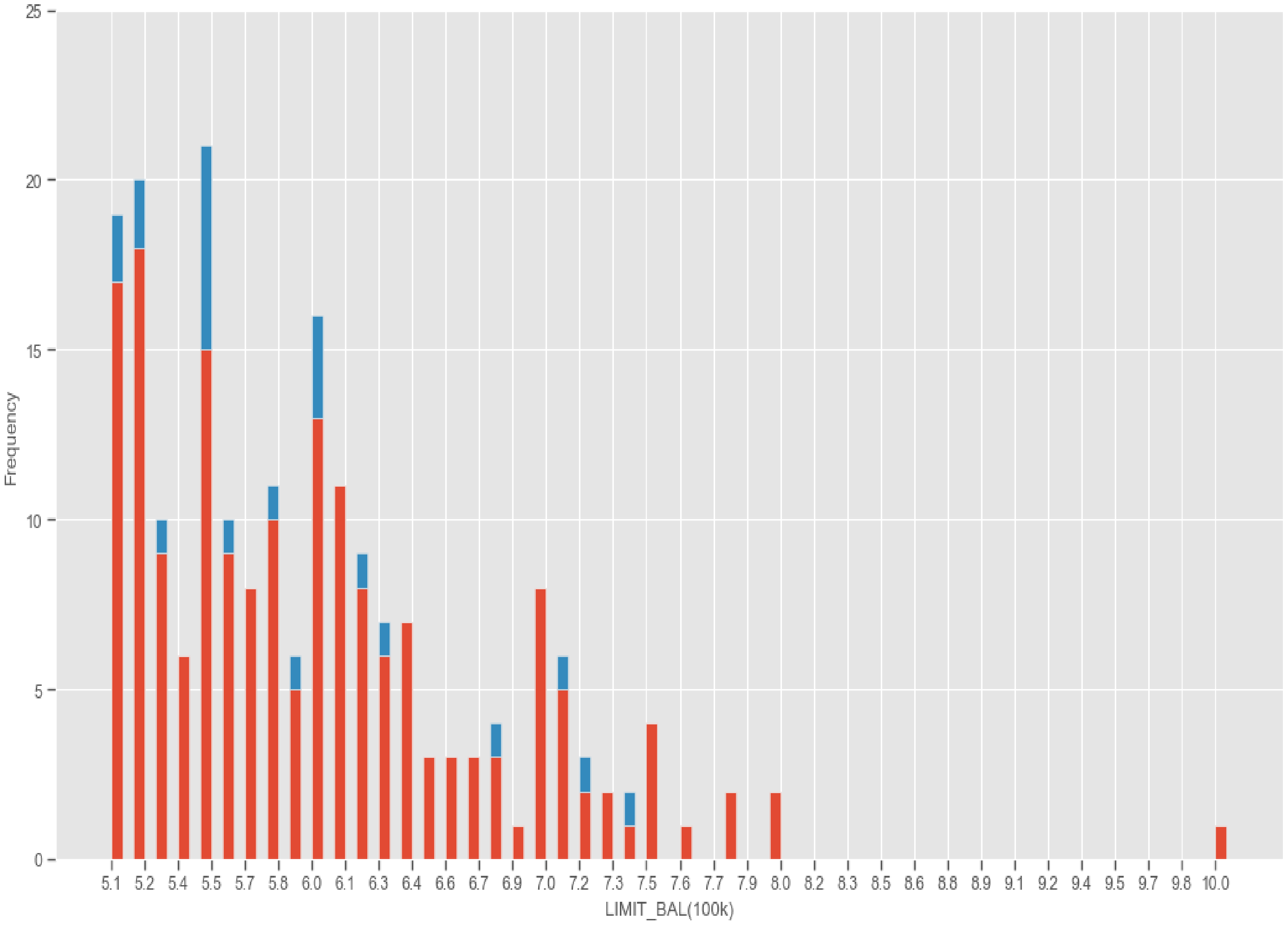


Data Exploration

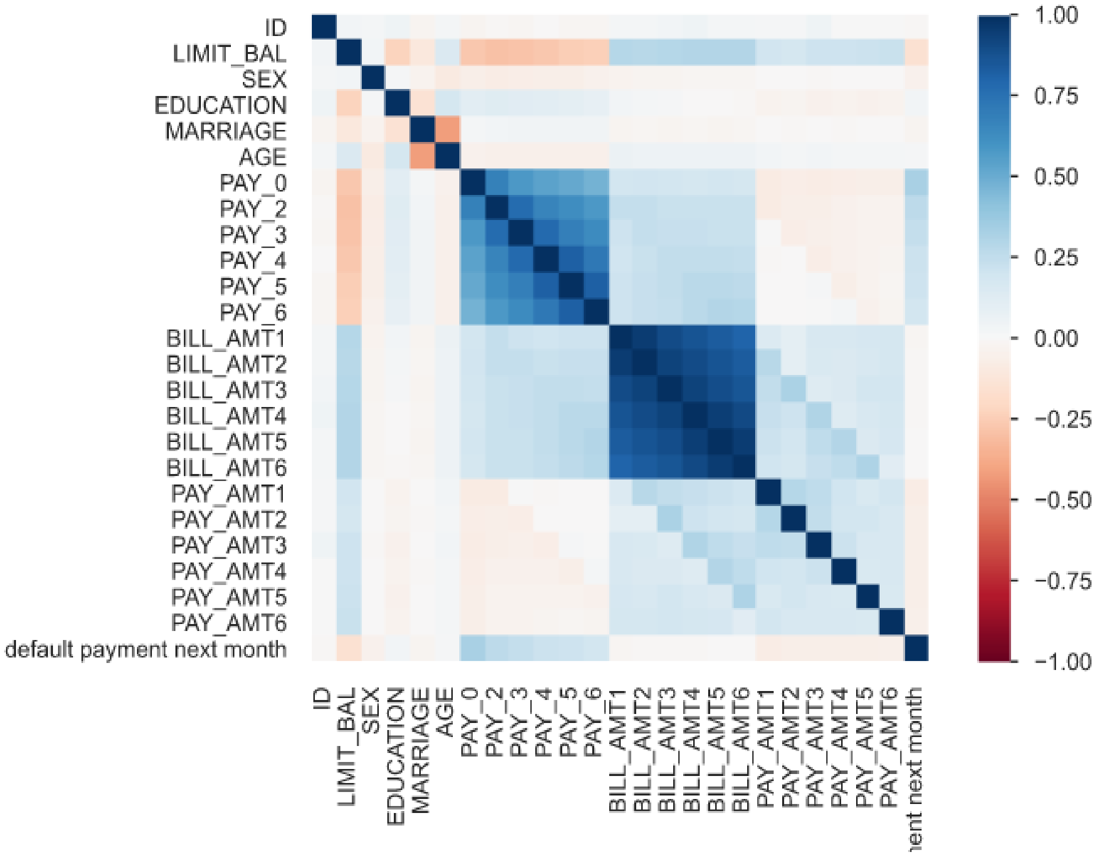


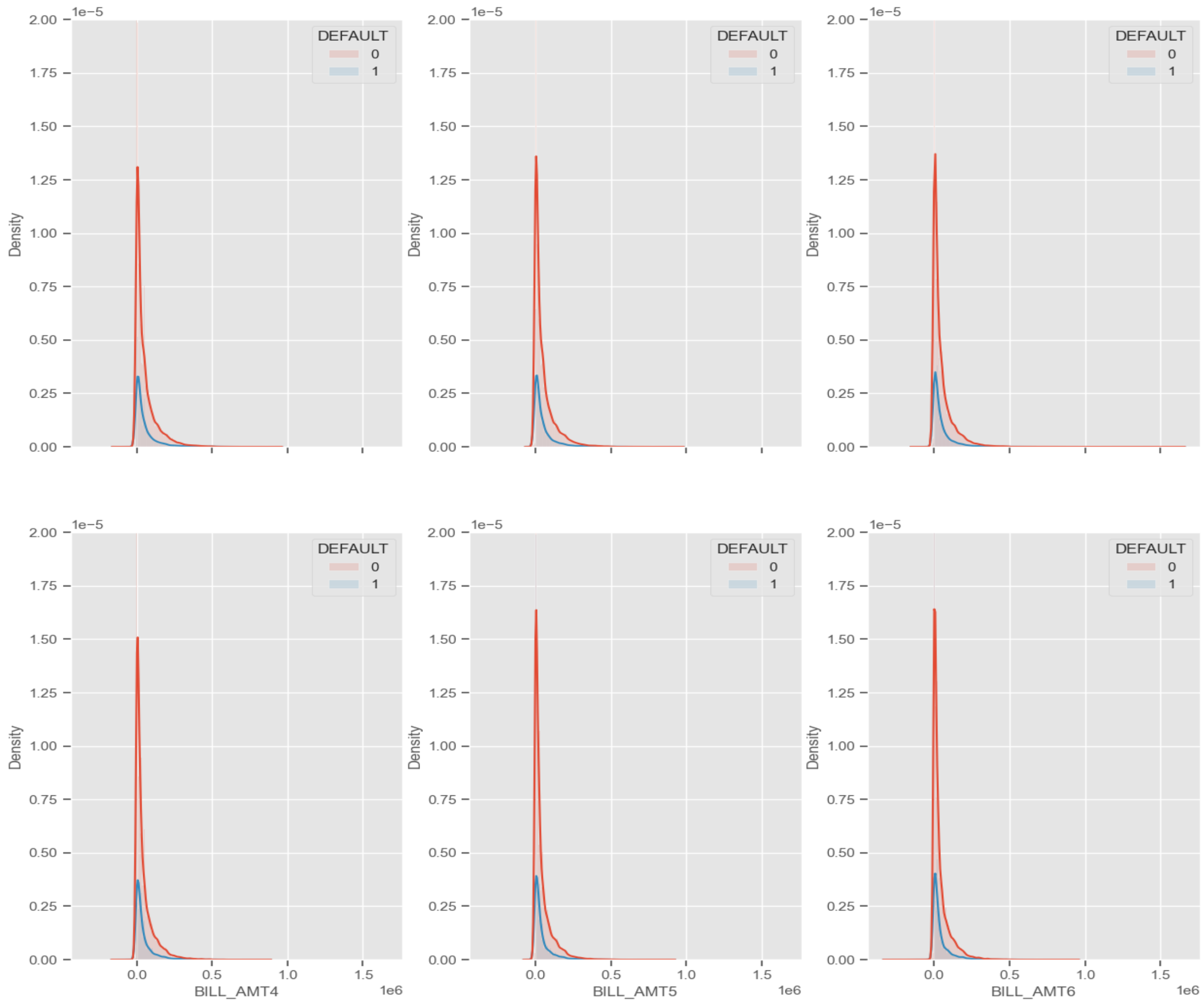
Data Exploration





Data Exploration

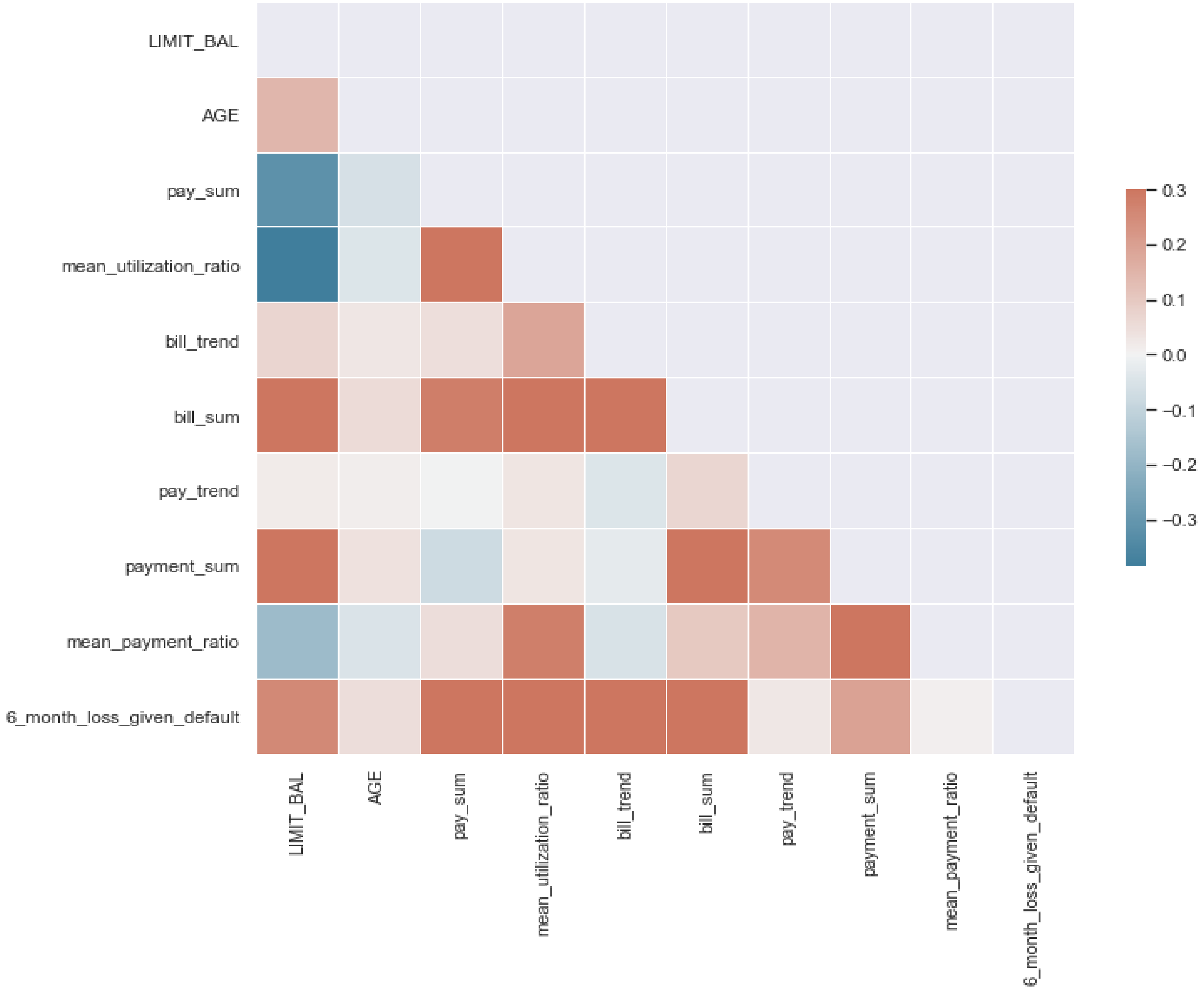




Feature Engineering

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. $payment\_sum = \sum_{i=1}^{6} PAY\_AMTi$

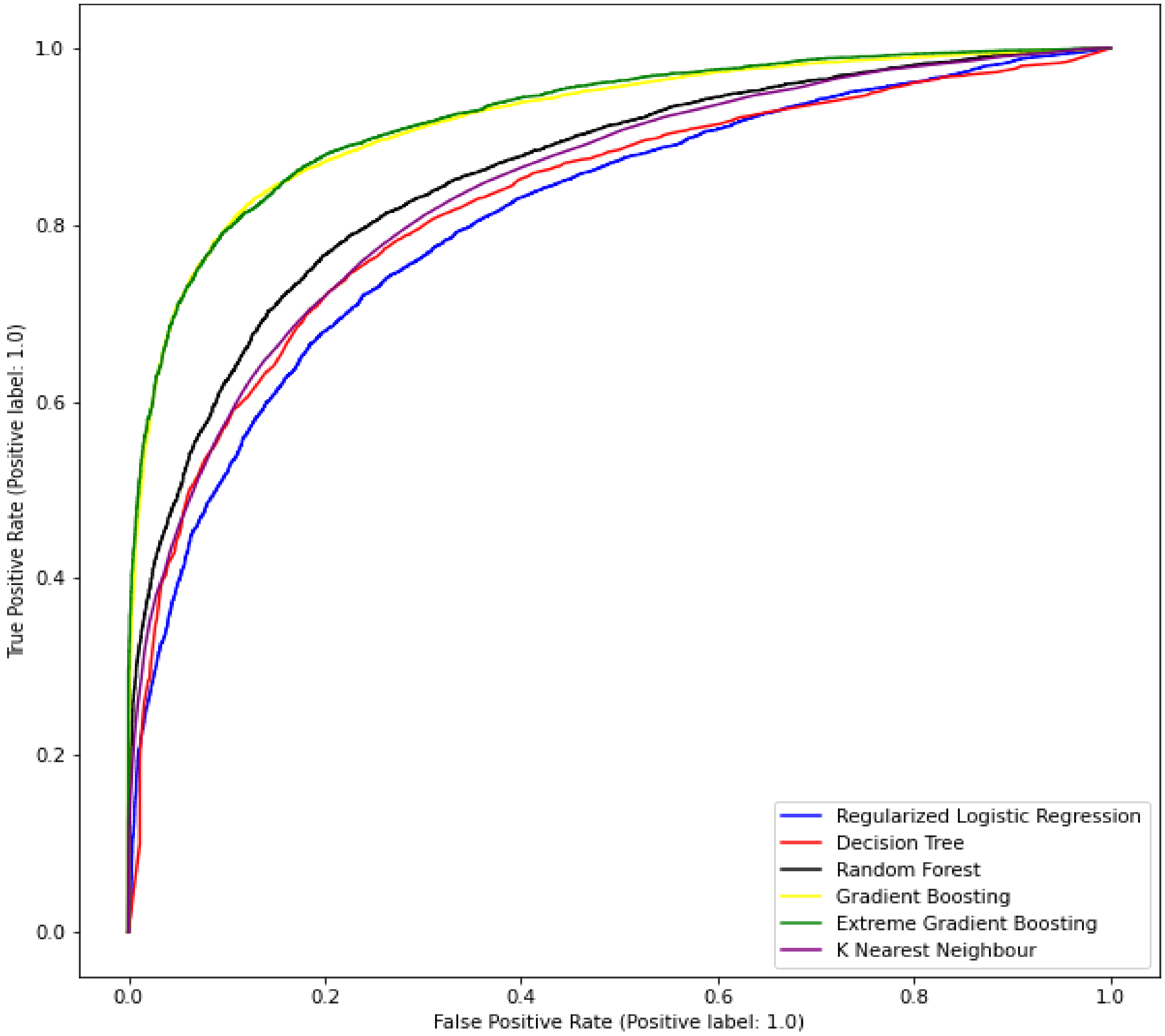
Feature Engineering



Modeling using the original data set

Algorithms	Hyper parameters	best estimate	accuracy	precision	recall	AUC	runtime	top 3 most important features
regularized logistic regression	C	0.01	0.739	0.747	0.725	0.808	7.26	PAY_1, MARRIAGE_2, SEX_2
decision tree	criterion, max_depth	gini, 10	0.757	0.787	0.707	0.808	26.07	PAY_1, PAY_2, MARRIAGE_2
random forest	n_estimators, max_depth	200, 9	0.782	0.814	0.732	0.858	182.25	PAY_1, PAY_2, SEX_2
gradient boosting	learning_rate, n_estimator, max_depth	250, 9, 0.25	0.844	0.863	0.819	0.917	1926.3	PAY_1, PAY_2, BILL_AMT1
Extreme gradient boosting	subsamples, n_estimators, max_depths, learning rate, gamma	0.8, 400, 10, 0.05, 0.3	0.846	0.867	0.817	0.923	418.62	PAY_2, PAY_1, EDUCATION_4
KNN	n_neighbors, p	50, 1	0.76	0.782	0.721	0.84	921.95	PAY_1, PAY_2, LIMIT_BAL

ROC curves comparison



Modeling using engineered features

Algorithms	Hyper parameters	best estimate	accuracy	precision	recall	AUC	top 3 most important features
regularized logistic regression	C	0.05	0.727	0.733	0.716	0.795	pay_sum, MARRIAGE_2, EDUCATION_3
decision tree	criterion, max_depth	gini, 10	0.756	0.77	0.73	0.814	pay_sum, payment_sum, MARRIAGE_2
random forest	n_estimators, max_depth	200, 9	0.773	0.784	0.754	0.851	pay_sum, payment_sum, MARRIAGE_2
gradient boosting	learning_rate, n_estimator, max_depth	250, 9, 0.25	0.827	0.829	0.825	0.901	pay_sum, payment_sum, pay_trend
Extreme gradient boosting	subsamples, n_estimators, max_depths, learning rate, gamma	0.8, 400, 10, 0.05, 0.3	0.826	0.829	0.82	0.901	pay_sum, MARRIAGE_2, EDUCATION_4
KNN	n_neighbors, p	1, 50	0.754	0.772	0.72	0.829	pay_sum, LIMIT_BAL

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, RF, GB, and KNN.

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.

- Features: We didn't have credit bureau data in this project. We also applied historical data, not the information recently.

Future work

Considering the popularity of deep learning model, it is worthwhile to explore its application in this data set.

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