Credit Score Alternative: Scoring Risk of Default using Transaction Data

Kuangyu Zou kzou@ucsd.edu
Haicheng Xu hax004@ucsd.edu
Xuewen Yang xuy001@ucsd.edu
Jessica Guzman jeguzman@ucsd.edu

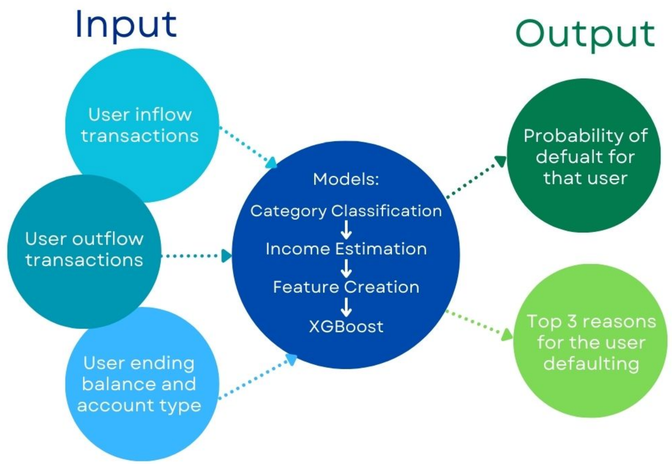
Mentors: Brian Duke brian.duke@prismdata.com, Berk Ustun berk@ucsd.edu

PrismData

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

Background & Introduction

The current methods used to predict default scores use credit history and repayment behavior. This paints an imperfect picture of a potential borrower because it ignores information embedded in a user's daily cash flow. We Develop a model that combines all features related to income, balance, and transaction categories to predict whether the customer will default their money to banking.



Data Cleaning & Summary

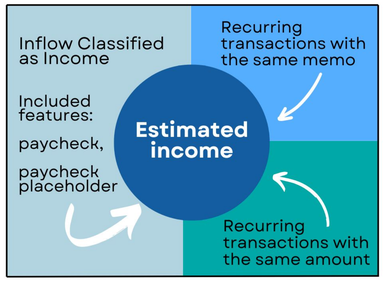
- **Features**: Consumer IDs, consumer account IDs, transaction information, transaction amount, transaction date, evaluation date, and balance.
- **Cleaning**: Blurring any phrases containing numerical digits to enhance confidentiality.

memo
ATM CHECK DEPOSIT 08/06 XXXX BAILEY RD CUYAHOG...
ATM CHECK DEPOSIT 09/10 XXXX BAILEY RD CUYAHOG...
Zelle payment from CODY CRANO CTZ0H8QOXXXX
ATM CHECK DEPOSIT ON 04/08 213 STEPHANIE, STE ...
ATM CHECK DEPOSIT ON 10/25 213 S STEPHANIE ST ...

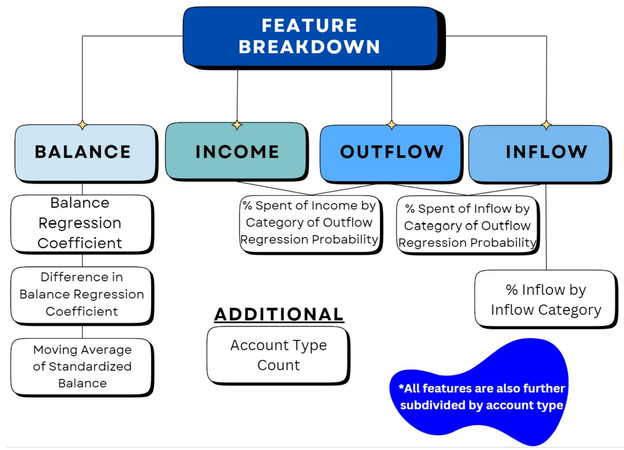
memo
ATM CHECK DEPOSIT X/X XXXX BAILEY RD CUYAHOGA ...
ATM CHECK DEPOSIT X/X XXXX BAILEY RD CUYAHOGA ...
Zelle payment from CODY CRANO X
ATM CHECK DEPOSIT ON X/X X STEPHANIE, STE X He...
ATM CHECK DEPOSIT ON X/X X S STEPHANIE ST STE ...

Income Estimation

One component we knew we wanted to include in order to assess risk of default was income. We believed people with steady income were more less likely to default, but since this measure is not monitored, we needed to estimate it. Our estimate depended on:



Feature Engineering



Feature Selection

Results

Table 1. Model Comparison

Metric		Logistic Regression	SVM	XGBoost
precision	0	0.82	0.81	**0.88**
	1	0.63	0.00	**0.59**
recall	0	0.99	1.00	**0.93**
	1	0.09	0.00	**0.44**
f1-score	0	0.90	0.90	**0.90**
	1	0.16	0.00	**0.50**
accuracy		0.82	0.81	**0.84**
auc		0.81	0.79	**0.87**

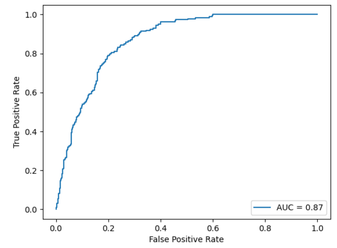


Figure 1. ROC Curve for XGBoost

The best performance was achieved when using a subset of 35 features with the XGBoost model. Accuracy: 83.72% , AUC: 0.87. In addition, the ROC curve gives us further insight for profits/loss incurred from the model. For a given point .2 to .8 on the x and y axes respectively, the model correctly accepts 80% of non-defaulted loans while incorrectly accepting 20% of defaulted loans.

We also found the most common reason codes for users predicting as defaulting, as seen below.

Table 2. Top 10 of Most Common Reasons in Percentage

Feature	Percentage (%)
CREDIT_CARD_PAYMENT_outflow_over_income	33.93
Predictions_cat_proba	26.43
checking_month7_EMA	18.29
checking_month4_EMA	18.26
CHECKING_balance_std_diff_regress_coeff	16.68
EXTERNAL_TRANSFER_inflow_over_income	16.27
EXTERNAL_TRANSFER_inflow_over_inflow	15.70
checking_month5_SMA	14.66
MISCELLANEOUS_inflow_over_income	14.02
SMALL_DOLLAR_ADVANCE_inflow_over_outflow	13.82

Limitations

The model's performance in predicting class 1 is suboptimal, with precision and recall values of 0.59 and 0.44, respectively. This is caused by the composition of the data: More than 80 percent of customers belong to class 0 (non-defaulters), while less than 20 percent are in class 1 (defaulters). Consequently, the model tends to prioritize optimizing precision and recall for class 0 at the expense of class 1.

Contributions Beyond

We leveraged transaction data from user accounts, creatively converted them into useful features, and trained a state-of-the-art model with those features. We also try to adhere to strict ethical standards, refraining from utilizing features that could lead to discrimination against protected classes. Overall, our model links a larger expanse of borrowers to lenders who can adequately assess their risk and therefore improves the capacity of the financial lending system.