Predicting Missed Payments from Credit Card Clients

By Juergen Kritschgau and Song Gao

Can we use credit card client payment history to predict missed payments?

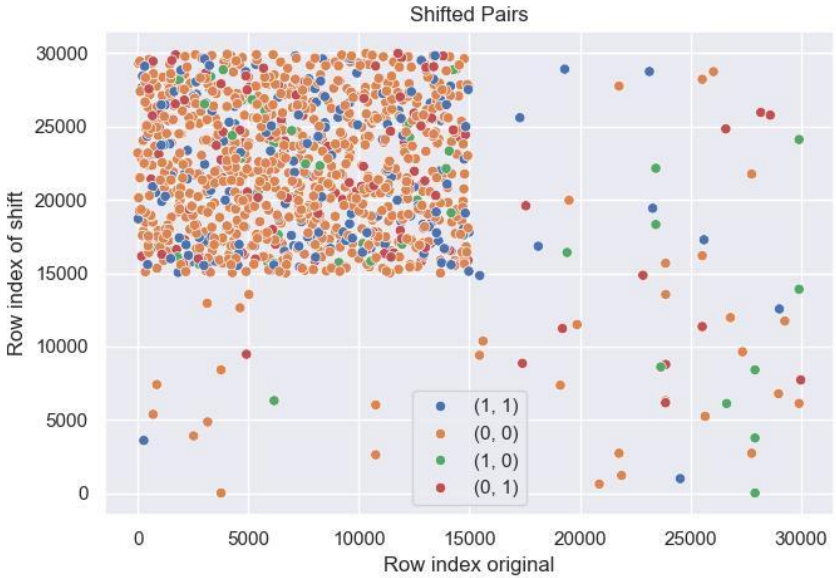
- Helpful for understanding credit worthiness and forecasting incoming payments for credit card issuers
- Target interventions to keep customers up to date on their payments
- Determine the value of defaulters to debt collectors

Dataset and Data Description

- We used the "Default of Credit Card Clients" data obtained from https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients. This data is publically available for use under the CC BY 4.0 license.
- There are 30000 instances where each row corresponds to the payment history and demographic information of a credit card client in Taiwan collected in 2005.
- Features included: bill amounts, payment amounts, repayment status, and basic demographic information.
- Using a 6 month window of payment information, can we predict if a client will make a payment next month?

Data Cleaning

- Duplicate rows
- Overlapping time windows
- Undocumented entries for
  - Marriage Status
  - Education
  - Payment Status

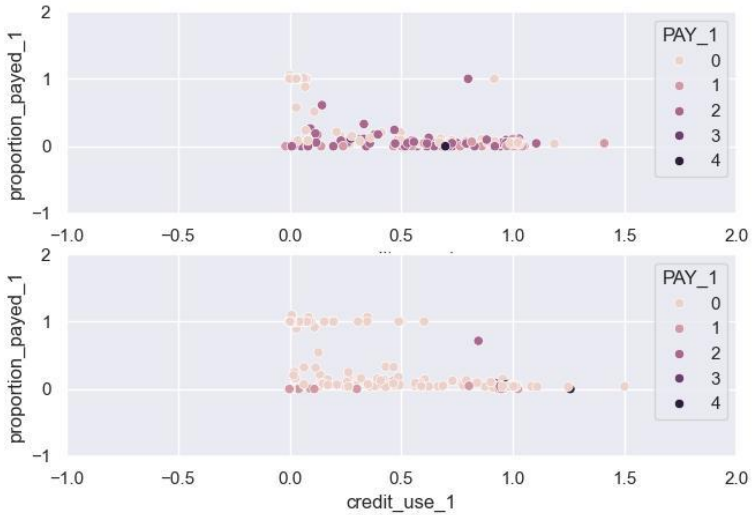


Feature Engineering

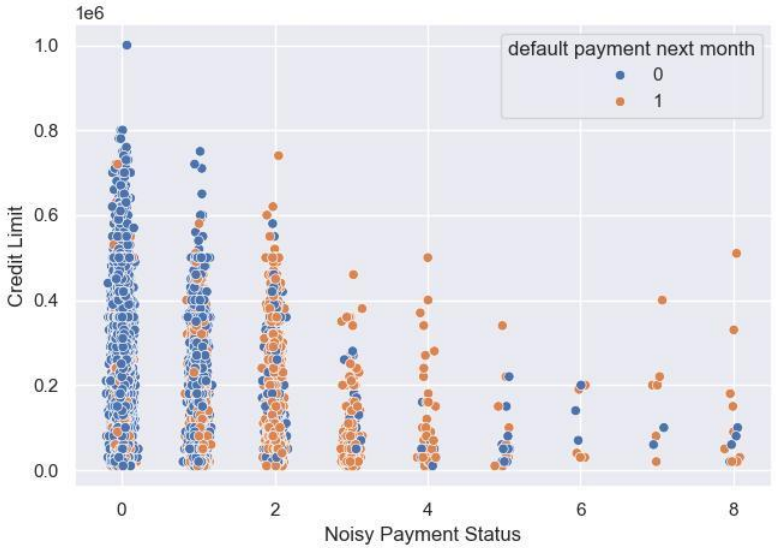
We want to add the following features:

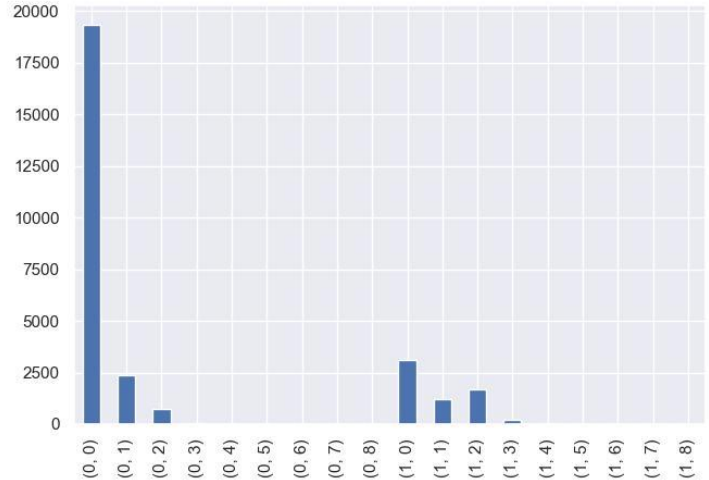
- - (Balance) / (Credit Limit) for each month
- - (Bill Payed) / (Bill Amount) for each month
  - note that this is relative to the previous months bill;
  - in case of division by 0 we replace with 1

The plot down samples to 200 points



Exploratory Analysis





Data Pipeline

Some models require comparable data scales (e.g. Support Vector Machines, KNN). Interaction variables may be helpful as well. Our pipelines are model specific.

- Scale data columns by their mean and standard deviation
- Create quadratic interaction variables
  - Due to (balance)/(credit limit), and (Bill Payed) / (Bill Amount) this creates duplicate columns
  - Drop duplicate columns
- Train Model

Proposed Models

- k-Nearest Neighbors (kNN) Classification with k = 5, 10, 15
- Logistic Regression with and without interaction
- Support Vector Classification (SVC)
- Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Gaussian Naive Bayes (GNB)
- Decision tree

Model Validation

- Metrics
  - Accuracy: percentage of correctly classified data points
  - Precision: percentage of accurate missed payment predictions TP/(TP+FP)
  - Recall: percentage of missed payments accurately predicted TP/(TP+FN)
- 5-Fold Cross Validation Split with Stratification
- 5-Fold Cross Validation Split with downsampled training data

Models performance

We measure the performance of our models by accuracy, precision and recall scores.

- SVC was found to be the best model, with the following validation scores:
  - **Accuracy: 82.06% Precision: 68.04% Recall: 33.76%**
- A similar performance was also achieved by LDA.
- The GNB method was also remarkable with its high recall score. Scores for GNB:
  - **Accuracy: 33.67%, Precision: 24.00%, Recall: 92.90%.**

Note imbalance class sizes: predicting that a client will make a payment has 79% accuracy

SVC performance on testing data was **Accuracy: 82.61%, Precision: 70.81%, Recall: 34.81%**

Further Research

- Consider the probability of default instead of Boolean values.
- Test on new datasets from other banks.
- Dive deeper into feature engineering.
- Try other machine learning models.

Thank You!

We would like to thank the Erdõs Institute for hosting the Data Science Bootcamp. In particular, we appreciate lectures and support from Steven Grubkin and Alec Clott.