Introduction

Our client is the world's largest peer-to-peer lending platform, providing loans and investment opportunities for individual and corporate customers. In recent years, the share of defaulted loans issued on the platform has increased significantly. We have been asked to help develop a model to support the underwriting process. The aim of the project is to increase detection of defaulted loans before the loan is issued/offered on the platform, so the model score can be used as an additional input in the application assessment process.

We propose using a Gaussian Naive Bayes classifier to predict default. The model trained on the full dataset performed as below:

- 82% of the loans that defaulted were correctly identified as being likely to default
- 71% of default statuses were predicted correctly
- 82% of loans predicted to default did default.

Definition of business metrics

We propose three business metrics to be monitored when this model is implemented:

- Accuracy: what proportion of loan statuses were predicted correctly?
- Precision: what proportion of loans that we predict will default, do default?
- Recall: of the loans that defaulted, what proportion did we identify correctly?

We have identified recall as the most important metric to optimize for in this case.

Methodology

The dataset provided by the client consists of 2.26 million loans issued on the platform between 2007 and 2018, with 150 recorded attributes for each loan.

We define a default client as one with loan_status taking on one of the following values:

- Charged off
- Default
- Does not meet the credit policy. Status: Charged Off
- Late (31-120 days)

A complementary definition of a non-default client is one with loan_status as one of the following values:

- Fully paid
- Does not meet the credit policy. Status: Fully paid

All other clients in the dataset are ones with an indeterminate loan status, and are removed from the data to reduce possible bias.

Data preparation

We take a number of steps to prepare the data before performing predictive modelling:

1. Log scaling of 13 columns with a skewed distribution to reduce bias in the normalization step. These columns primarily relate to income.
2. Removal of columns providing little information, either due to missing values or high granularity.
4. Extraction of information from categorical columns, either by conversion to numerical form (eg. credit grade) or setting a boolean flag.
5. Normalization of numerical columns, performed after train/test split.
6. Removal of columns whose values are only added after the loan is issued.

Possible sources of data bias

We have identified four possible sources of data bias:

1. Joint applications make up 5% of the dataset, but there may be differences in borrower profile between these and individual applications. We have removed joint applications from the dataset pending further investigation.
2. The dataset consists only of issued loans, providing us no samples of loans that were denied before issuance. We have reduced this problem by training on an undersampled dataset with an equal balance of classes.
3. We cannot assign a class to loans still in the process of being paid off – these may default later or be paid off successfully. We remove these loans from the dataset.
4. We did not remove points which appeared to be outliers (eg. unusually high income) as these may be the result of skewed income distributions.

Instead, we scaled these and related columns using log scaling.

Modelling approach

We prepared data with a 90:10 train:test split, undersampled to have balanced classes. We tested four models: Logistic Regression, Decision Tree, and Random Forest, and Gaussian Naive Bayes. Gaussian Naive Bayes was chosen based on metric results in test (see appendix).

Recommendations for further work

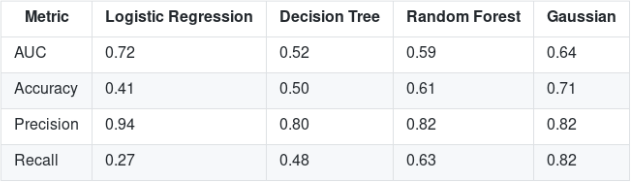
1. Perform analysis of issued loans in 2007-2008, 2009-2014, 2015-2018 to identify possible demographic changes affecting default rates. Default rates in the first and last groups are consistently >20%, but < 20% in the middle group.
2. Assess if there is significant difference between individual and joint borrowers to further inform model training.
3. Collaborate with the client to further incorporate credit scoring and sub-scoring into model.

Conclusion

We have successfully developed a model which correctly identifies 82% of loans that later default. Identification of these loans could allow, in the best-case scenario, a reduction of up to $150m in charged-off loan value per quarter (based on a 10% rate of charged-off value in 2017 Q1 and an issuance value of $1.9b in 2019 Q1: https://www.lendingclub.com/info/demand-and-credit-profile.action).

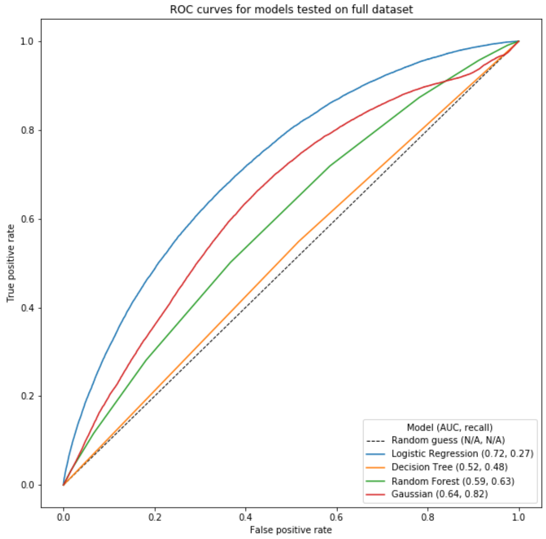
Appendix

Figure 1: Metric scores for the four models trained and tested on the full dataset.



Metric	Logistic Regression	Decision Tree	Random Forest	Gaussian
AUC	0.72	0.52	0.59	0.64
Accuracy	0.41	0.50	0.61	0.71
Precision	0.94	0.80	0.82	0.82
Recall	0.27	0.48	0.63	0.82

Figure 2: ROC curves with AUC and recall metric values for the four models trained and tested on the full dataset.


ROC curves for models tested on full dataset
True positive rate
False positive rate
Model (AUC, recall)
Random guess (N/A, N/A)
Logistic Regression (0.72, 0.27)
Decision Tree (0.52, 0.48)
Random Forest (0.59, 0.63)
Gaussian (0.64, 0.82)