Data Source

Data

The dataset that we will use is a subset of data from a public competition in kaggle, there are 307511 observations and 122 variables in the dataset. The data points are for each loan application and its applicant. The variables range from applicant demographics such as number of children to income and information about the loan/application itself.

Data Manipulations

Missing Values

A significant portion of the dataset's features contained NULL values. We did a filtering that removed columns where more than 40% NULL values existed and then removed observations containing NA value. Motivation of removal was that if many values were missing then the feature is not important to the dataset.

Outliers

For the dataset's numerical features, there were found to be many outliers that severely skewed the distribution of the dataset. Instead of removing records with outliers, capping outliers was performed.

Feature Selection (Filter Methods)

Performed Pearson pairwise correlation analysis for independent variables and removed one from a pair where a high correlation (>.7 or <-.7) was found.
There were 20 similar variables called “Flag Document n”. Along with strong assumptions that these variables would be irrelevant, they were removed after having found extremely low correlation with the target variable.

Dummy and Categorical Variables

Most of our dataset consisted of categorical variables, so in order to numerically represent their values, they were either converted to n-1 dummy variables (when the categorical feature is not ordinal) or numerically scaled (when the categorical feature is ordinal).

Table Joins

We have two datasets, one is current loan information and the other is previous application information. We want to use both current applications and previous ones, so we join both applications and calculate the times that one client was approved or rejected.

Resampling

The data was quite unbalanced with far more 0 than 1. We want to focus more on 1 in the model to successfully predict the potential default. So resampling is necessary. After trying undersampling and oversampling, we decided to use undersampling since we have quite enough observations. Different ratios of undersampling were used, for which was evaluated for best performance during model building and development.

Data Assumptions

Insignificant Variables

As per our group's interpretation of the features, many of them seemed unnecessary to our analysis. Information about the home for which the loan applicants intended to purchase, such as the number of building floors, were removed. Other binary features were assumed to be unimportant and therefore removed such as if the applicant provided a phone number.

Outliers

With features such as income and loan amount, although there were many extreme values, we avoided the removal of them because they were assumed to be important. We deemed that removal would result in loss of important information representative of high income applicants.

Analysis Summary

Modeling

We have built five classifications models:

1. Logistic regression (with threshold = 0.5)
2. Classification tree
3. K-nearest neighbor
4. Random forest (with threshold = 0.5)
5. Naive bayes classification

We compared this model using PPV (positive predictive value) and NPV (negative predictive values), which basically are the models' class 0 and class 1 prediction accuracies: PPV indicates the ratio of predicted 1 class that are truly 1 and NPV is the same for class 0.

The most important metric for us was NPV. We picked the best model based on the highest NPV, which will mean that it has the lowest false negative amounts. Since wrongly predicting a class 0 (a right payer) as class 1 (a defaulting payer) is costlier for the bank, we tried to minimize this error. Moreover, with the given dataset, the best result for PPV was only 30% with a drop in NPV of 10%, therefore we decided to minimize the number of false negatives.

Other Important Analysis (descriptive, predictive, prescriptive)

Descriptive

In order to understand all the features in our dataset, we visualized the distribution of our continuous variables via histograms and boxplots. For categorical variables, we visualized them using tables and bar plots in order to get an idea of how many unique categories there were. Basic summary EDA of variable means, maximums, minimums, and standard deviations were calculated to further aid of understanding of the dataset.

Predictive

As stated before, the main objective of our analysis was to predict which customer will have problems paying back the loan. As we worked more with the data, our predictive models and analysis would work best as a first screening.
Since our model's predictions on class 0 are quite accurate, the bank can be quite confident in identifying class 0 customers.

Prescriptive

If the model predicts a class 1, the person would need another background check. For the prescriptive analysis, we would recommend the company to have a second screening method for class 1 predictions. Since the accuracy of our model is quite low, it would probably mean that the variables given might not be useful to classify a defaulting payer or the model might be too complex with too many variables.
One suggestion on our end is to get the credit score from another source and try to create a simple model with only the different credit scores, which seem the most relevant variables.

Analysis Conclusions

Further Considerations

A key takeaway from our analysis is that to properly model the data, it may be beneficial to not treat the data like a “one shot fits” all. Rather with the intuition being that different groups of people may have different prediction models, more classification based on clients features may improve model performance. Moreover, because it is difficult to place all applicants under the same umbrella, this would not be the best approach in capturing the wide variation in applicant behavior and circumstances. The best approach could perhaps be to look into the best segmentation criteria for the applicant base such as demographics before conducting a separate model for each.

Relevant Variables and Interpretations

- Ext_Source_2: this is the credit score of the customer from an external source: With a higher score, the odd of defaulting payments is way lower (exp coef of ext_source_2 is 0.0985)
- The status of the previous applications
- The contract type: it is more likely that someone pays late in cash loans than revolving loans
- Days_employed: since it shows a stable income
- Education_type: the higher the education level, the lower the odds of late payments Owning a car lowers the odds of defaulting payments.

Resampling

With this dataset, we had difficulty in finding the best resampling strategy. By naively using sampling methods on an imbalance dataset, it could generate bad results on a class accuracy, since the imbalance might be reflected in the real world but not on our trained model.

Performance Metrics

The model performs well on class 0, therefore if it predicts 0, then it is highly likely that the applicant will pay in time. However, the model doesn’t perform well on class 1. It is possible to use our model as a first check to divide optimal customers from not sure clients and then do another background check for class 1 predicted customers. The model performs well in predicting class 0, with over 94% accuracy, however not so well in class 1.

Data Governance

Our business recommendations are based on how the appropriate stakeholders such as financial institutions and credit unions could further leverage the dataset to benefit their specific use cases. In the case of banks and how they could further enhance their predictive models using the dataset, they should have a clear and detailed history of the applicant's previous applications. Having those records highly impact the decisions on approving the loan, therefore an organized applicants’ history can help the bank to decide on an applicant request.

For the secondary background check, the bank should ask for additional variables not used in the models and in the dataset. For example, the company could ask for another external source to rate the credit score of the applicant, since the credit score was the most impactful in the model.