STAKEHOLDER REPORT

Identifying delinquencies in the auto loan industry

Summary

What are the patterns in registered loans data and how can we predict whether a customer default on the loan?

Alexander Prip, Claus Mørkbak Højrup, Mark Daniel von Kelaita and Martin Jæger Nielsen

Introduction:

When the financial crisis struck in 2007 - 2008, lenders and other financial institutions experienced significant losses. These institutions have lent a great amount of money on variable loan terms - the so called 'Subprime loans. Combined with the fact that the background checks on their clients were far from thorough, this was a disaster just waiting to happen, just as it did.

The outcome of this were a tightening of already established regulations and lenders making the required actions to try to identify whether a borrower would default on their loans. Yet in March 2019, The Federal Reserve reported that the number of borrowers with auto loans that were more than 90-days delinquent in the fourth quarter had reached a peak of 7 million in total. This was an increase of 1.5 million compared to the previous quarter and had led us to ask ourselves the question if lenders and financial institutions have taken the proper measures to perform proper background checks and whether their methods are suitable. Based on this we are intrigued to explore if statistical methods and machine learning can be applied in order to predict, if a borrower will make their payments on time and not default and whether we can identify any patterns in the delinquencies.

Initial description of the dataframe:

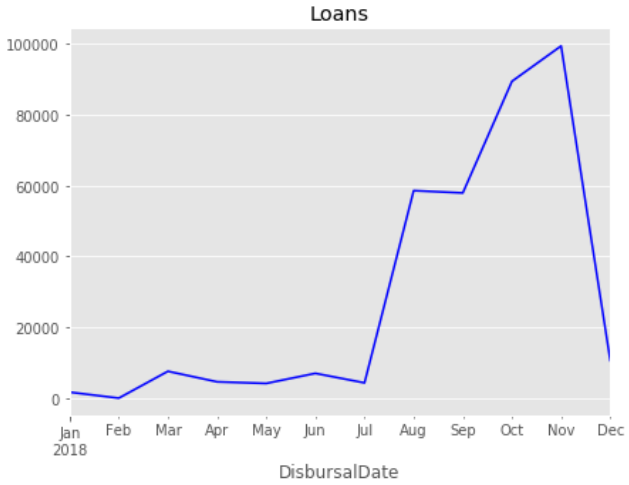
The dataframe consists of 345.546 unique loans, which are loans that have been issued to a borrower**.** During the screening process, information has been collected on the borrower (demographic data; age, pincode/zipcode, documentations provided etc.), the type of loan (disbursal details, asset value, LV ratio etc.), and the issuing branch (branch score, active accounts, loan status, and history). If possible, the branch has estimated a credit/risk score which ranges from 300 to 900. The different features/columns in our dataframe amounts to 36 features after initial feature selection.

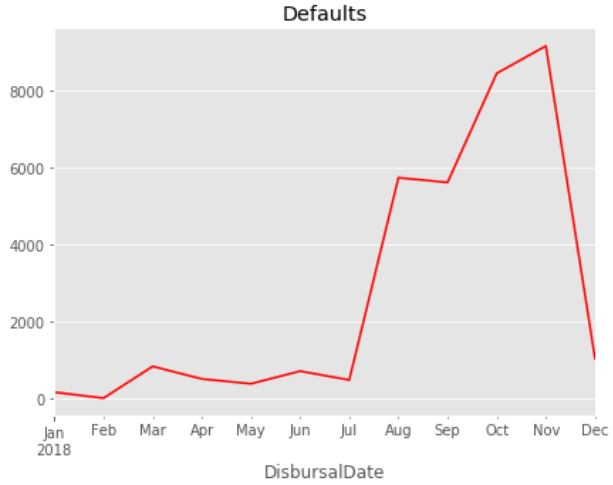
Further investigation:

From the dataframe some initial exploration has been performed. As depicted below in the tables and visualisations there is a total of 82 branches and 3398 employees. By manipulating with the dataframe we can derive that the branches perform very differently in assessing whether a borrower will pay their instalments and interest on their loan. Out of the total issued loans 4.718 are delinquent equal to or above two months. This translates to roughly 1,4 percent. However, when we explore the total number of delinquencies (incl. delinquencies after one month) it's much higher.

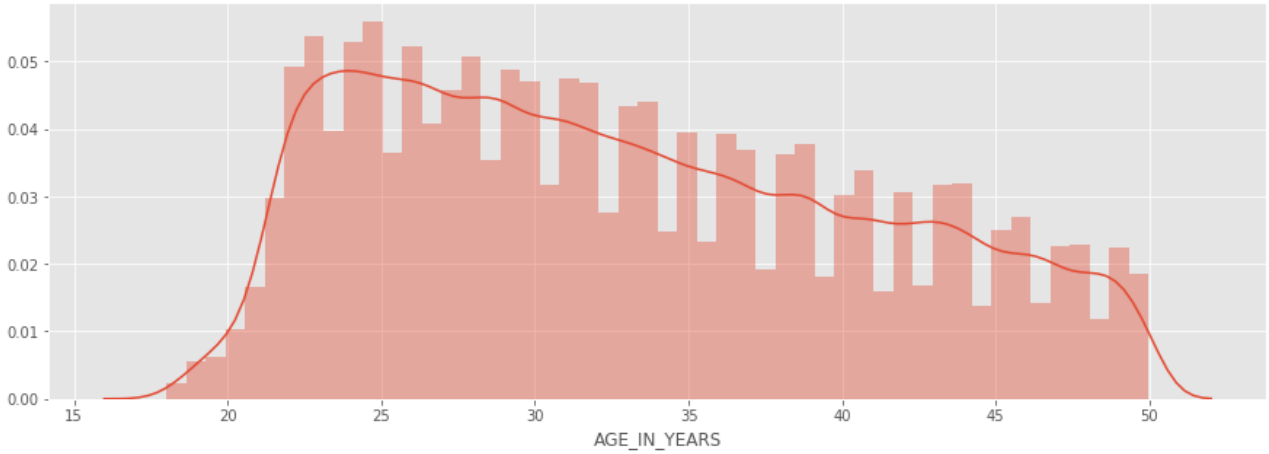
As mentioned, a fair number of borrowers are delinquent after just one month and although this is an issue, it can be argued that this could be a 'one time' occurrence and that the borrower in the coming period will

pay their loans in time. None the less we stress that this is an issue, however we want to explore the share of delinquencies that occur two months or above in a row, as this behaviour only strengthens that the borrower may completely default on their auto loan. The graph below depicts the issued loans and delinquencies equal to or after two months. Comparing the graphs there aren't a lot of fluctuations between them. Thus, the trend delinquencies follow the trend of issued loans.

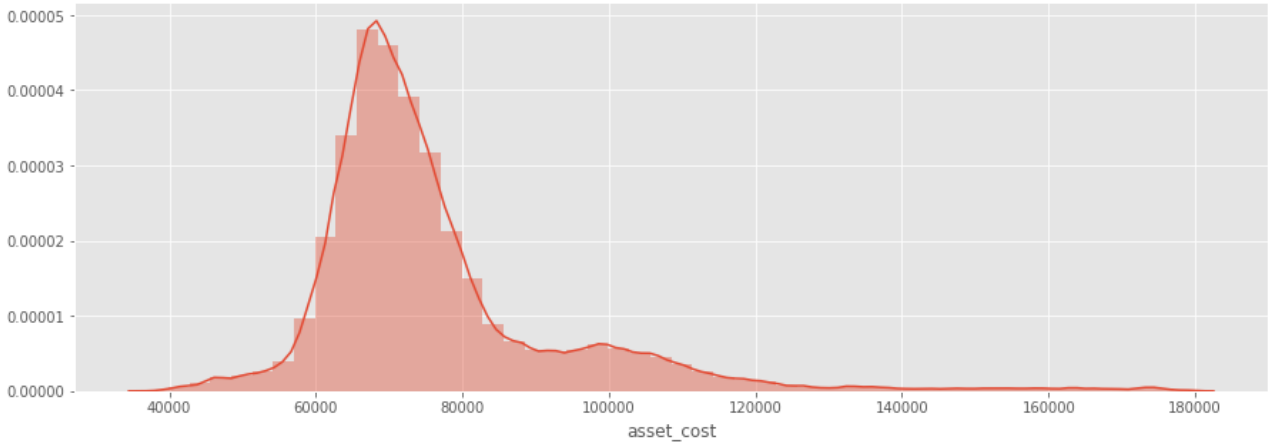




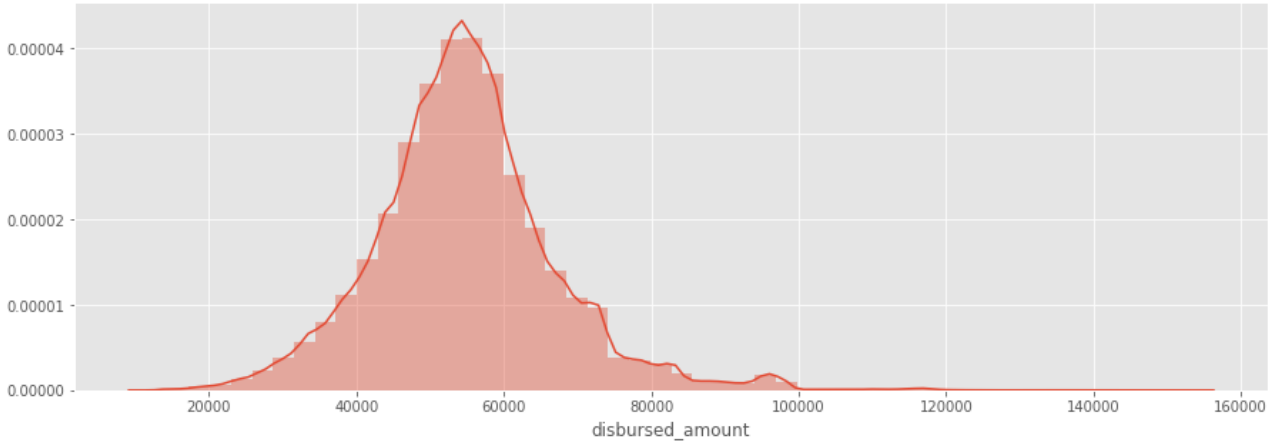
Further exploration shows the age of the borrowers whom the loan was issued to. The trend shows that the major share of borrowers is in the age group 20 to 30 approximately peaking at the age of 23 and then slowly declining until the age of 49 where a sharper down going trend occurs.



Another interesting element below is the asset distribution which displays the amount each car was worth at the point of sale. Having filtered extreme outliers we find what could be argued as a normal distribution for the asset cost. Before the outliers were removed, we had a mean of 76.498 USD. This can be compared to the disbursed amount to see if there's a relationship.



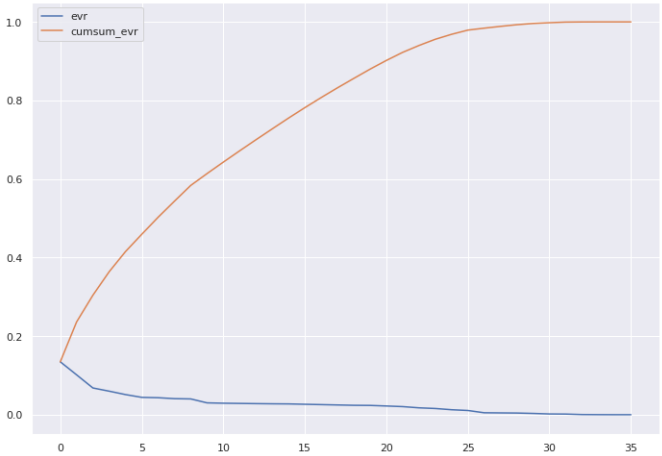
Comparing the two graphs we find a similar relationship between asset_cost and disbursed_amount. However, the mean of the disbursed amount is 54.836 USD which tells us that the borrowers at average puts a down payment of roughly 20.000 USD, which seems like a fair amount considering the mean of the total purchase. It's important to note that the removed outliers from asset_cost and disbursed_amount would have affected the graphs and dataframe, but in order to provide the most useful visualisation, these have been removed. The biggest outlier for asset_cost had a value of 1.628.992 USD and though it's possible for a car to have that price we consider it to better to remove it.



Patterns in the dataframe:

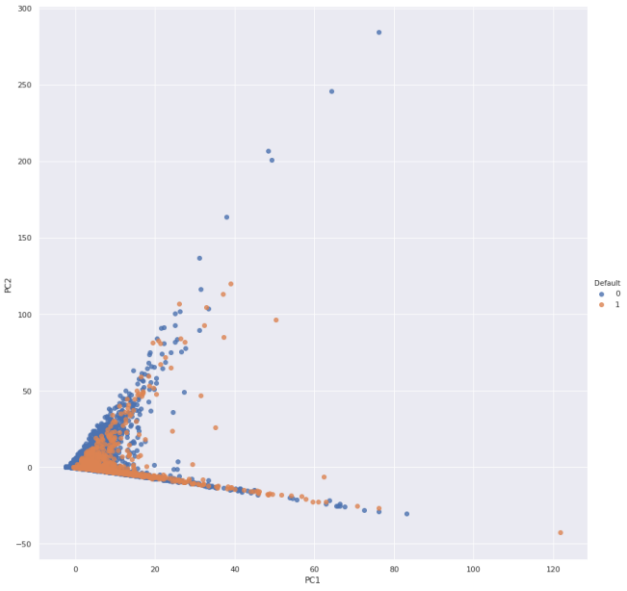
As mentioned in the beginning we're dealing with what could be argued to a large amount of features. For instance, we collect information on all the different types of ID's the borrower would be able to provide. Wouldn't it be enough to register if the customer provided valid ID or not? Perhaps, but such a distinction would be subjective on our part, and we would not know the value of the information that we removed.

Instead methods like dimensionality reduction, which this report won´t detail, can compress the information into fewer dimensions. Illustrated on the plot below it shows that all 36 components aren't required in order to retain 100 percent of our information. In fact, we could reduce the dimensions to 30 and still retain 100 percent of the value in the dataset. As a rule of thumb 90 percent of the data´s value should be retained. Looking at the graph, we see that 20 components will get us a little less than 90%. Thus, we choose 23 components.



With our information compressed into 23 components we can do further analysis to identify which parameters are the most important ones. From the heatmap, which can be found in the notebook the highlighted colors indicate the parameters which are the most influential/dominating in distinguishing which features are important in defining patterns in the data.

In order to get a further understanding of whether there are any patterns between the issued loans and occurring delinquencies a cluster visualization is shown below. Given that the two components plotted falls approximately in the same pattern gives an indication that the delinquencies which occurs are random. Therefore, no specific brand or employee can be held responsible for borrowers not being able make their monthly payments.



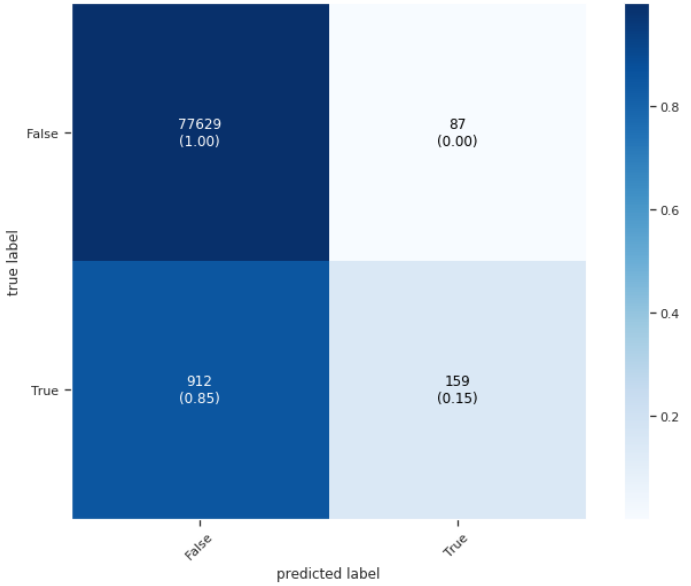
Had the graph had areas where the orange or blue were structured in a non-arbitrary way where one color was much more dominating than the other, it would have been an indication that specific branches, employees or other features had a performance that were unlike the majority.

Classification:

In order to predict whether borrowers will be unable to make their monthly instalments, some classification reports will be needed. Using multiple methods, we model this to get an overview and basis to compare

similarities in the output of the models as well as judge if the results seems reasonable. The three models used are; Logistic Regression, Decision Tree, and XG Boost.

Comparing the test scores provided by each model, we can derive that they perform with similar accuracy with a difference of roughly less than a percent. However, the model best a predicting was XG Boost with an accuracy of 98,7 percent and for illustration the confusion matrix to illustrate how accurately it predicts is inserted below. The upper left corner shows us the True Negatives. eg. How many loans were classified as not default and how many were in fact not default. The lower right corner tells us the True Positive. eg. How many loans were classified as default and how many were in fact default. The upper left and the lower right corner is where we want as many observations as possible – as this tells us who will be able to make their monthly instalment and who will not. In the upper right and lower left corner, we can see, where the model is incorrect in its prediction. We would want a model, that minimizes the observations in these areas of the confusion matrix.



Looking specifically at the confusion matrix to the right 98,5 percent of our observations falls in the True negatives corner, thus most of the borrowers make their monthly instalments. The model predicts with success that 0,2 percent of the borrowers are unable to make their instalment. However according to the model 1,2 percent of the borrowers predicted being to make their instalment are not able to. In addition, 0,1 percent of the borrowers are classified as being unable to make their payments though they are.

Conclusion:

It's up for discussion whether an accuracy of 98,7 percent is enough for a model to be used in this context. Considering that the dataframe had outliers with a value in excess of 1.000.000 USD, it could have severe consequences if a borrower that had been issued loan worth that amount was unable to make a monthly instalment or in worst case default. Additionally, with the interest rates being as low as they have been for a while, it could be argued that the margins on the loan market are thin, and thus the need for highly accurate models are strengthened.