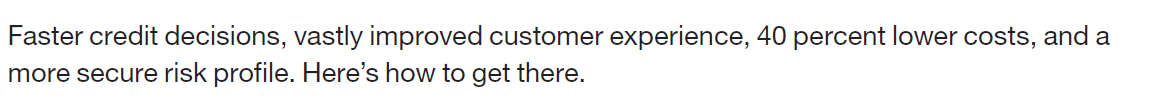
Background:

A:

Slide 1

Digital lending is the next hot thing in commercial banking. A recent McKinsey report summarized digital lending as following:



Slide 2

Most of the digital lending implement some machine learning based PD model. Compared to traditional scoring models, they incorporate macro-economic variables, and their outputs admit a probability interpretation. Compared to structured models, their inputs are readily available, and their forms structures are more flexible.

B:

Slide 3

The biggest players in digital lending are Peer-to-peer firms like Lending Club. These players typically do not take risk themselves; they initiate loans, and sell them to investors. Due to their ability to customize interest rate and their lighter operating structure, P2P lending has grown significantly over the years. In 2018, about 35% of personal loans in the US are originated by P2P lending companies.

Slide 4

In this project we do an EDA on past Lending Club Data, and build a default classifier. The training data contains loans issued between 2007 and 2011, and the test in 2011. All of the loans have concluded, meaning they have either been defaulted on, or fully paid.

Our most extensive effort is concentrated in data visualization and pattern extraction. For modeling, we attempted several methods to overcome the imbalance in dataset, but the results weren’t ideal. We will discuss possible improvements.

Data Cleaning

A

Slide 7-9:

One major feature of the data cleaning process is that we wrapped our procedures in pipelines. We do this so that first future users can more easily replicate our process, and second, such modularization gives us easier control over what to do and what not.

Slide 10:

So what did we do in data cleaning?

Inside the dataset, we first grouped the features into 4 baskets: Loan Condition, Borrower Financial Strength, Credit Situation and Payment Information on the Loans. This is so that we can think about the more than 40 variables in a more systemic way.

We then removed empty columns and columns without variations, then formatted the variables.

B:

Slide 11:

External data

One of the variables in the data set is the three digit zip code.

At first we thought geographical position of the loan application might give some information about the quality of the application through maybe the neighborhood, but this interpretation of the geographical position of the application is already contained in the annual income variable.

So we started thinking more on what to infer from the geographical position and started looking at the census data of the 2006-2010 American Community Survey estimates.

The census data has data points for the median house value, median gross rent, per capita income and unemployment statistics. However the granularity of the data is not at the same level as three digit zip code. The census data has a block of size around 3 thousand people which is called a census tract.

So we needed to map census tracts to three digit zip codes. Missouri Center for Data Science provides a ttool to map zip codes to census tracts.

However some three digit zip codes have more than one census tracts in it. There fore we grouped the census tract data calculating average metrics such as average per capita income, unemployment rate, and although it is not threoretically correct, we calculated average median house value and gross rent and mapped these values to individual loan applications.

A:

EDA

We can build a predictor with 100% accuracy with just one line – simply checking whether the total principle received is less than the loan amount. This is because conditioned on loan having concluded, that’s the exact thing that determines default!

This is apparently too good to be true. The problem is that our dataset is conditioned on loan having concluded, and for many variables their values conditioned on loan having concluded introduces correlation that won’t exist when the loan is still on going.

Slide 12:

These variables include: payment data on the loan and last credit pulled date. We think lending club does active credit check on on-going loans. From the histogram we can see every month some loans mature, and some loans go through their last credit check. Then at 3 year, a batch of loans mature, and go through their last credit check. This is exactly the sort of behavior if lending club does a regular credit check.

If that is true, then last credit pulled date reflects the how long Lending Club monitors a borrower’s credit situation. For on-going loans, that period would be the same as the loan’s lifespan, and contains no particular information about default.

B:

Slide 15

Then there are variables that are based on PD, and won’t be available in prediction time. They include Loan Grades, interest rate, installment

Slide 16

What’s to note is that the ratio funded by investor is actually an exogenous variable. When investors give their money to Lending Club, they explicitly indicate the range of loan parameters they are interested in. Lending Club, being an intermediary rather than an actual risk taker, will pass as much of the loans to investors as possible. This means that each loan's percentage funded by investor is determined by how much investor money is available for this type of loan, which reflects market’s overall confidence in this risk basket. We therefore keep it.

A:

Slide:

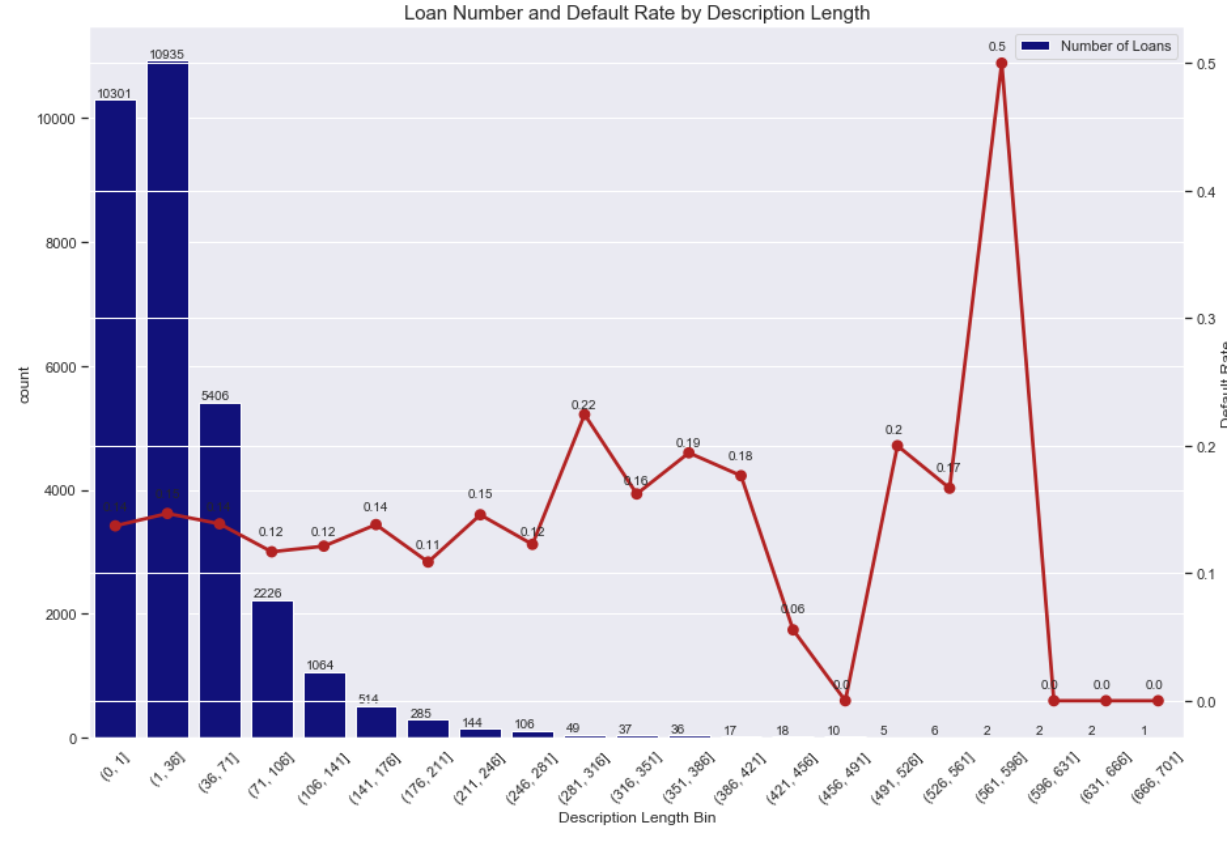
For the remainder of the variables, we proceed basket by basket. From each basket, we first select a couple of variables that are too important to exclude, then incrementally test the inclusion of new variables. We plot the newly added variable against the existing ones, and if there’s a visible correlation, we do a hypothesis testing using LRT.

Slide EDA – Term

For example for the term variable, we discover that 3-year loans have a 10% default proportion, while 5-year loans have 22%. So term has some correlation with default. But how much of that can be explained by loan size? Even though the average loan size for a 3-year is smaller than a 5-year loan, hypothesis testing shows the coefficient for term indicator is significant. This means term still has some explanatory power not captured by loan size. We include this.

**EDA** Description length

We read on a paper Caldieraro et. al. (2018) find that there is information on not the content of the loan descriptions of lending club applications but rather on the length or the absence of description. We tested this result ourselves to see if there is any significant variability in the default percentages for different description lengths.



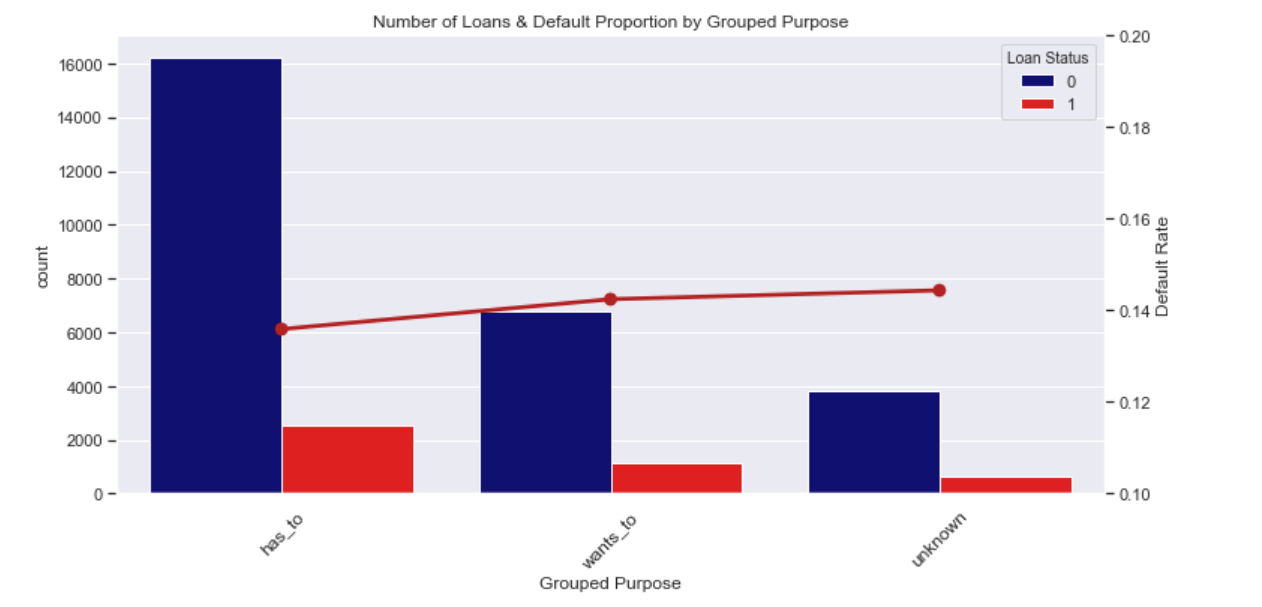
We see that even though there is ver few data points at the far end of the description length distribution, there still is about 2-3% difference in average default percentage. We consider this to be significant fo we included this variable as well.

**Loan Purpose**

The other variable is the loan purpose. An initial look at the default rates and number of loans at each category shows great variability on the default rates in each category. We wanted to further analyze this by grouping the loan purposes in three groups: has\_to loans, want\_to loans and other loans.

However there does not seem to be a significance with respect to the nature of the purpose.

Therefore we leave the purpose data variables as is, and do one hot encoding to transport into numerical variables.



A:

Slide: EDA – Income Outlier

Most people in the US get their first line of credit at around 19. We add that to credit history to infer the subjects age.

We see people under 35 claiming to own more than 1 million, and still renting a house? We remove them.

Slide EDA Income verification:

Countrary to common belief, both source verified and unverified incomes are lower than the verified group in distribution, and it is the verified group that has the highest default rate. We also observe borrowers with a verifiable income tend to apply for a bigger loan. A possible explanation is that borrowers with a verified income are more optimistic in their chance of getting approved, and will thus apply for a higher amount, which is risker.

We add interaction terms between income and verification status to reflect such dependency. This says the same income with different verification status only impact the default probability differently.

EDA Home ownership

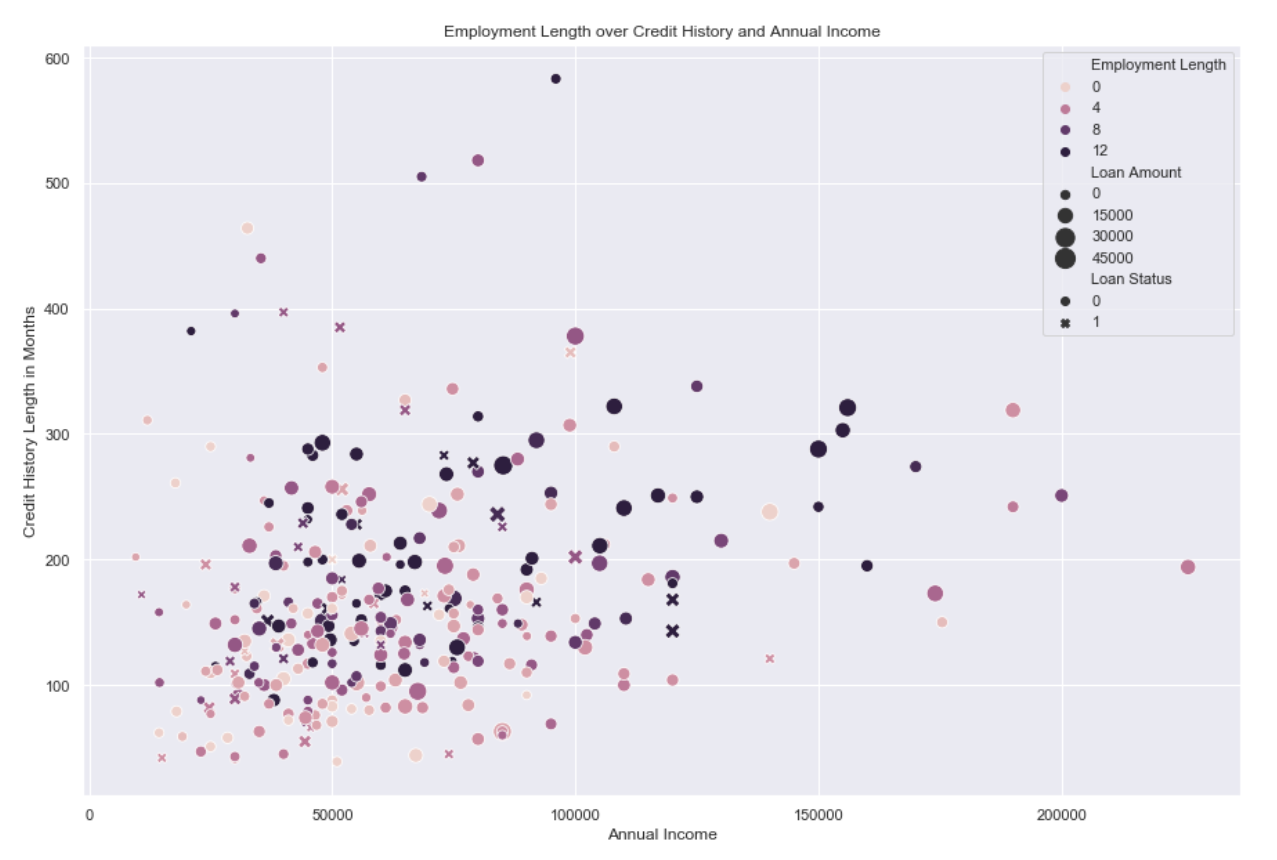
We also used indicator trick to reflect the property’s value’s impact on default in different home ownership status. If you own a house, it’s your house value that matter; if you rent a house, it’s the rent that matters.

**Employment Length**

-We transform employment length into numerical values between 0-10

-Around 800 missing data is filled randomly with values between 0-10.

Deeper color shows longer employment. Longer employment length tend to be on the upper right corner of the graph, implying higher annual income. But at the same time the points are bigger, meanin that the loan amounts are also large for longer employment lengths.



Also when we look at the average default rates for the different employment lengths, we again see a 2-3 percen variance. Therefore we are also including them in the model

EDA States

Lending Club's biggest presence is in costal states: California, New York, Florida and Texas account for more tha 46% of its entire loan pool. Yet in some states where embracement of technology is strong, its presence are pretty weak: in MA it only has 2 loans, and in Washington 710.

Even though there are fluctuations in default rate between states, we think including 47 indicators might add too many parameters in to the model. We have also tried to capture geographic information using region-specific macro-economic variables. Thus we won't include State in the actual prediction.

EDA Credit

Number of Total Accounts – Deragotory public records

Both of the variables show significance to the average default rate. We reason that number of open accounts implies the applicant being in the credit system, maintaining credit accounts. This gives additional information to the credit history of the applicant.

Similarly number of deraggotory public records also provides information regarding the history of the applicants ability to fulfill liabilities.

EDA Correlation:

No pair of the numerical features have a correlation greater than 0.5 (other than Rent and House Value, which is expected and acceptable, because by implementing the indicator trick we are using them on different parts of the dataset). This means our EDA has successfully ruled out repeated features, and captured different facets of the dataset!

Modeling

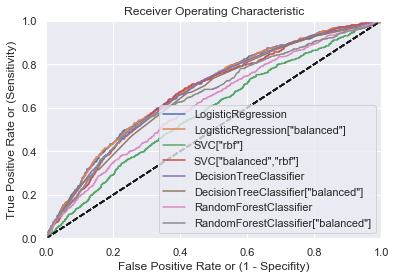
To get a sense of how different models performing, we selected 4 different prediction models for our initial test.

The models are logistic regression, support vector machines, decision tree and random forest models.

We the models to our prediction models without any cross validation, to check which models perform better.

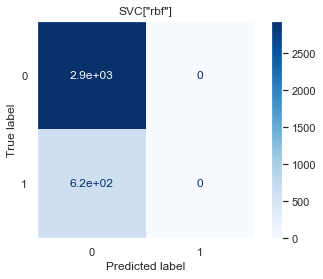
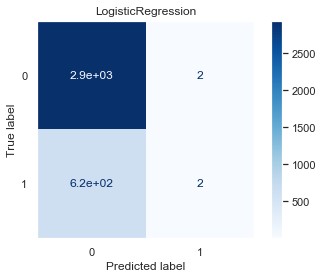
For each of the model, we implemented both the unbalanced method, that does not do any transformation to our training data set and we also implemented the same models with the same default parameters.

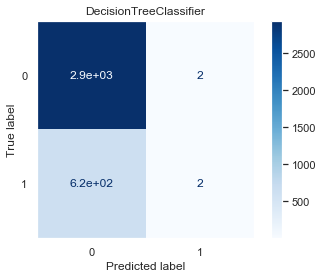
When we plot the respective ROC curves we see that almost all of the models have a similar roc curve

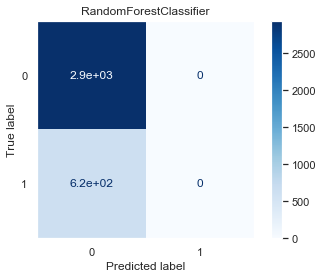


We then continued to plot the confusion matrices for the models using the test data.

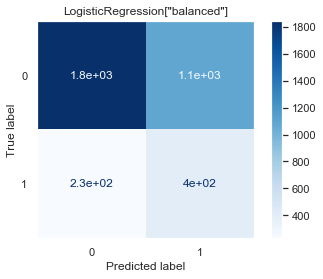
Now if we look at the confusion matrices we see that the models that are not balancing out the training data almost exclusively behaves like a trivial classifier.

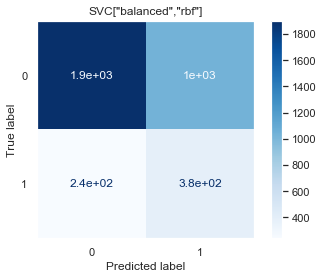


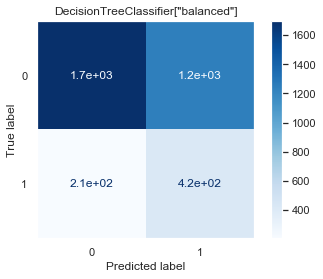


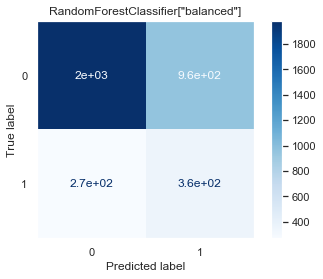


And for the models that are balancing out the data have a more balanced predictions, yet the performance of these predictions are obviously very poor.







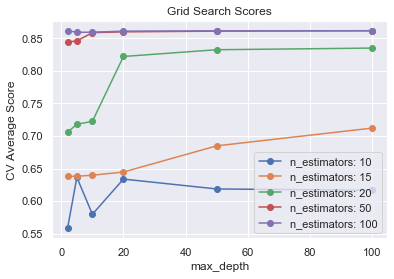


For that reason we selected two candidate models SVm and random forest with data set balancing parameters and moved on to optimize the model parameters.

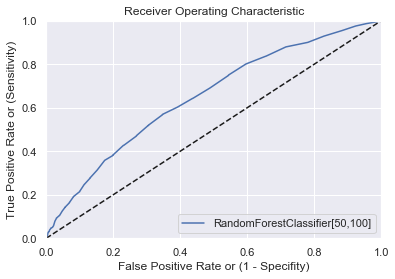
The first model is the Random Forest model.

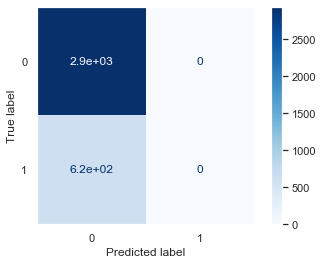
The grid search algorithm is applied to optimize the number of estimators and max\_depth parameters.

The optimization of the gridsearch algorithm is done over the mean accuracy of models for the kfold cross validation sets.



Once we optimized the parameters, we checked the roc curve of the models. We see a similar result for the ROC curve, compared to the initial model set results. The result does not seem to have improved at all.



However when we look at the confusion matrix, we see that the model behavior changed, and that Random forest again behaves like a trivial classifier.

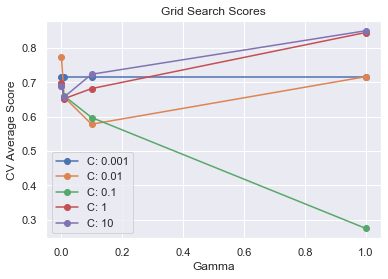
The accuracy of the model predictions is 0.8245860230143138

However, this model never predics a defaulting loan and acts completelt as a trivial classifier. Given that a false positive prediction should be desired over a false negative prediction, in our view this is a very poor fit for the data.

We conduct the same analysis for the SVM model with the balancing parameter. This time the grid search optimization is done over the gamma and regularization or the penalty parameters.

Again this grid search is optimizing the parameters over the mean accuracy of the k fold cross validation sets.

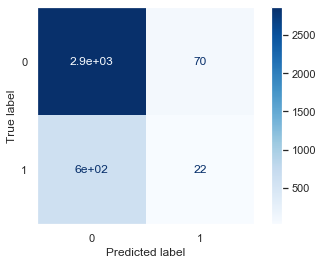
SVM grid search



The best model is obtained with gamma=1 and c=10;

The ROC curve is again showing a similar result without any improvements.

And the confusion matrix also shows that the SVM model, even though it is balancing the data acts as a trivial classifier.



Accuracy = 0.8111142295818131