|  |  |
| --- | --- |
| P2P Lending Default Prediction | |
|  | |
| February 2020Machine Learning II | R.Chawla, M.Jepson, M.Salhotra, A. Yildiz |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | IntroductionPeer-to-peer (P2P) lending arose from the movement to democratize the economy by utilizing the internet and new technologies. The key innovation being the removal of banks as the middleman in the borrower/lender transactions. This new process benefits both lenders and borrowers by allowing them to share the cut that would have otherwise gone to the middleman but what is going to replace the due diligence and risk management that was previously performed by the bank?Our project aims to see if this due diligence can be replaced by a borrower questionnaire, or in other words, can default on a P2P loan be predicted based on information about the borrower and the loan.In parallel we will review the work done by X. Ma, K. Sha and D. Wanxg in “Prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning” to see if their process and findings are replicable and consistent with the results we discover. | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data Being Used | |  |
|  | LendingClub is the largest P2P lending facilitator in the US and makes their loan data publicly available (subject to opening an account) at <http://bit.ly/2v0mU43>.  The dataset is very large in terms of both sample size and variable size. The dependent (y) variable is “loan\_status” which takes 8 values, 6 whilst the loan is outstanding and 2 once completed/defaulted. A default occurs when loan\_status is “charged off” and the no default has occurred when status is set to “Fully Paid”. Time Series and Filtering We noted in Ma *et al.*’s paper there was a time series element to this paper that is proposed to be have been caused by initial setup changes and the 2008 financial crisis.  After analyzing the loans based on issue date we found that the default rate was increasing slightly with time, contrary to the findings of Ma *et al.* It is possible that they were considering outstanding loans are not defaulted which would have reduced the rate for recent loans.  We can explain the increasing rate after 2014 to be survival bias as the loan lengths are 3 and 5 years. | Figure 1: The default rate over time from 2008 to 2018  Due to the size of the dataset we will use a single year’s data for training and testing our model. 2014 allows us to avoid survival bias as all the loans originated in 2014 will have completed by now (2020). |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data Clean Up Before creating a test-train split we needed to identify the data types in our data set as well as removing post-hoc  variables and unusable columns such unique ID’s and url's.  Data columns that are updated throughout the life of the loan such as “last fico score” and “interest received to date” were removed as they are not available at loan origination.  Categorical variables were encoded for random forest input with null values encoded as their own categories. Variable selection will be performed with random forest which can automatically handle variable transformation so label encoding can be completed without worrying about ordering or keeping ratios.  The information contained within null/missing values will be considered more closely if they are chosen by variable selection. Variable Selection A random forest with size 1000 was ran to identify variables with predictive value: Boosting Results Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. | Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.  Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.  Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since. |  |
|  | Image | |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lorem Ipsum is simply dummy text of the printing and typesetting industry.  Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. |  |  |