Project Proposal

Mengmei Chen, Yuze Liu, Boya Yan

March 28, 2016

1 Problem to solve

As Lending Club, a peer-to-peer lending company, becomes the world's largest online marketplace for borrowers and investors, we are interested in finding the best machine learning algorithm to predict loan defaults and help investors increase their return on investments.

2 Data and preprocessing

1. We will use all the existing data from the Lending Club website https://www.lendingclub.com/info/download-data.action. The data files include complete loan information for all individual loans issued from 2007 to 2015. Each row represents an individual loan and each column represents one aspect of the loan. If the "loan_status" is "Fully Paid", we label the loan as good (1). If the "loan_status" is "Charged Off", we label the loan as bad (-1).

2. We found out that the data is seriously imbalanced (data labeled "Fully Paid" is far more than data labeled "Charged Off"), so we decided to balance the data for future SVM learning. The way to balance the data is to make random copies of "Charged off" data.

3. Some of the features are empty for most or all of the data, so we remove those features.

4. Also, we consider some of the features as irrelevant to our prediction or very hard to make use of. Hence, we plan to remove them. Those features include the two ids, url, last payment date, last payment month, next payment day, employer title, issue date, payment plan, purpose, title, initial list status and payment amount.

5. We will transform the categorical features into boolean features by adding one boolean feature for each category and labeling it 1 if the feature is in that category and 0 otherwise.

6. There is a feature called “desc” that records how the borrowers described the loans. We plan to use SVM and/or TF-IDF to discover crucial words that distinguish the bad loans from the good loans.

7. Other than existing factors, we might add a few additional macrofactor features like unemployment rate by state.

8. We will normalize all the features to [0, 1].

9. Finally, we will split the data into a training set (80%) and a test set (20%).

3 Algorithms

1. We plan to use SVM to find a binary classifier for the data to predict whether a borrower is likely to pay back his loan. We will also try multiple kernels, see whether using L1 or L2 norm gives more accurate results, and find the best parameters for the models.

2. We plan to try out Random Forest to see if we can get a better result than SVM.

3. We might also use some algorithms that is to be taught in future classes.

4. Other than all the supervised-learning algorithms above, we will also try to use some unsupervised-learning algorithms such as clustering, to see if we can find out more information about the loans.

4 Evaluation

1. Because our data will be processed to be balanced, our baseline accuracy score is around 50%.

2. We will use cross validation to test the error rates of our chosen hypotheses. We will also calculate sensitivities, specificities, and PPVs of different hypotheses. Among them, specificity is the most important, because our ultimate goal is to help investors detect, among all the bad loans, as many bad loans as possible. In addition, we will give a detailed analysis on why a specific classifier performs badly and how changing parameters will affect a specific classifier.

3. Based on the previous analysis, we will choose the best hypothesis for the data and use the test data to evaluate our final hypothesis.

4. We also plan to calculate the increase in return on investment if investors had not chosen the loans that we predict to be bad.

5. Moreover, we will use clustering on all those ”Fully Paid” data. We hope that by using clustering, we could have a better understanding of those individuals who have successfully paid back their loans and suggest more efficient advertising strategies to the Lending Club.

5 Conclusion

From doing the project, we expect to learn how to preprocess the data more efficiently and have a deeper understanding of the various machine learning algorithms.

Table 1: Timeline

Week	Date	To Do
Week 1	Mar 28 - Apr 3	Data Preprocessing
Week 2	Apr 4 - Apr 10	Data Preprocessing
Week 3	Apr 11 - Apr 17	Algorithms and Evaluation
Week 4	Apr 18 - Apr 24	Algorithms and Evaluation
Week 5	Apr 25 - May 1	Final project write-up