CS 475/675 Project Proposal

Siwei Bian, Anna Burkholder, Ruixuan(Emily) Zhai, Chloe Warren
sbian1, aburkho7, rzhai2, cwarre19

Abstract

The abstract should consist of two sentences describing the motivation for your project and your proposed methods.

1 Project choice

Choose either a **methods** or **applications** project, and a subarea from the below table.

□**Applications**

□Genomics data □Healthcare data □Text data □Image data ✓Finance data

□**Methods**

□Fairness in ML □Interpretable ML □Graphical Models □Robust ML □Privacy in ML

2 Introduction

Explain the problem and why it is important. Discuss your motivation for pursuing this problem. If necessary, give some background on published work in this related area. Clearly state what the input and output is. Be very explicit: "The input to our algorithm is an English sentence, image, etc.. We then use a SVM, neural network, linear regression, etc. to predict COVID case count, text sentiment, etc.." This is very important since different teams have different inputs/outputs spanning different application domains. Being explicit about this makes it easier for readers. 1-2 paragraphs.

Our task for this project is to develop a machine learning algorithm that predicts people's loan eligibility. The input to our algorithm will be a series of personal information including gender, marriage status, number of dependents, education level, income level, loan amount etc. We then use a Random Forests Classifier to predict whether this person will be eligible for the loan. Our motivation for pursuing this problem is to explore a more efficient, transparent and consistent procedure in evaluating loan eligibility for batches of people as opposed to using human judgement on each individual loan application case, which is susceptible to human bias and error.

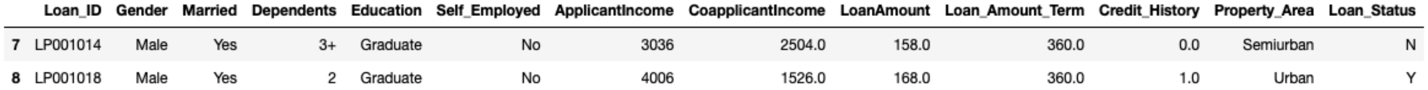
3 Dataset and Features

Describe your dataset(s): how many training/validation/test examples do you have? What pre-processing did you do? What about normalization or data augmentation? What is the resolution of your images? How is your time-series data discretized? Include a citation for the dataset(s) you are using. You should also talk about the features you used. If you extracted features using Fourier transforms, word2vec, PCA, etc. make sure to say so. If you have space, include one or two examples of your data in the report (e.g. include an image, a slice of a time-series, etc.). 1-2 paragraphs.

The dataset we are using contains 614 unique loan candidates. However, some of these candidates, who each have a unique loan ID, are missing data in one or more features in the given dataset. To rectify this, rather than augment the data to account for these gaps, we have decided to omit altogether the candidates with missing data in at least one feature column. After performing this pre-process we are left with 480 unique loan ID's, each with a full row of data.
We will then split this data into training, validation, and test data portions. We will assign 70 percent of the data, i.e. 336 candidates, to the training data. With the remaining 30 percent we will denote half of it (i.e. 15 percent of the entire set) as validation, and the other half as test data.
Our initial dataset contains 11 features, excluding each candidate's loan ID and their loan status (a binary feature which we hope to predict). Also, since our dataset is small, we need to further limit the number of features to make our model focus on more important aspects. Thus, features like gender of the candidate might be eliminated during this process. As we build our model it will become more

CS475 Machine Learning 2020, (LateX template borrowed from NeurIPS.)

clear which features are informative to the model, and which are not.
Our dataset comes from kaggle.com, and we will preview it below: `https://www.kaggle.com/burak3ergun/loan-data-set`



	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
7	LP001014	Male	Yes	3+	Graduate	No	3036	2504.0	158.0	360.0	0.0	Semiurban	N
8	LP001018	Male	Yes	2	Graduate	No	4006	1526.0	168.0	360.0	1.0	Urban	Y

4 Methods

Describe the methods you plan to use: what is your model's hypothesis class? your loss function? your optimization approach? Include enough information to demonstrate your understanding of the methods. You plan to use something not covered in class, explain it in 1-2 sentences, and provide a citation. 1-2 paragraphs.

Loan acceptances/rejections are unbalanced in reality due to the unbalanced nature of the American socio-economic structure and separation of classes. There are factors in play beyond what we have access to that impacts an individual's ability to secure a housing loan. We want to use random forests since it should play nicely with this issue, as explained in detail in this paper (`https://www.sciencedirect.com/science/article/pii/S1877050919320277#cebibsec1`).

With the use of random forest comes many decisions to make along the way, such as number of trees, max depth, min samples required to split or reach a leaf, and whether bootstrap samples will be used. For some of these parameters, such as our loss function, we have a choice going in to our project. We plan to use random forest with the default loss function specified in the Python3 sklearn package, "gini" to look at impurity as a way to measure the quality of a split. We also know we plan to use bootstrapping since we are working with a very small dataset. For other parameters, such as weighting or number of trees, we plan to alter as needed to see how it changes our model. The default optimization approach using gini is CART. (scikit-learn uses an optimised version of the CART algorithm; however, scikit-learn implementation does not support categorical variables for now.) We can encode our prediction to be 0 and 1 to use this. We plan to start with equal weighting and modify weightings as an attempt to gain higher accuracy if needed. The weighting scheme will be changed based on correlation of a feature with loan status if equal weighting doesn't give us the results we want. We believe this approach to our parameters will yield a richer understanding of the nuances of both our problem, and the classifier itself.

5 Deliverables

These are ordered by how important they are to the project and how thoroughly you have thought them through. You should be confident that your "must accomplish" deliverables are achievable; one or two should be completed by the time you turn in your Nov 19 progress report.

5.1 Must accomplish

1. Separation of loan acceptance/rejection.
2. At least 65 percent accuracy.
3. Go through a full implementation of the problem using Random Forest Classifier.

5.2 Expect to accomplish

1. At least 70 percent accuracy
2. Mixture of qualitative and quantitative data model inputs
3. Assess accuracy with at least two types of error measurement.

5.3 Would like to accomplish

1. At least 90 percent accuracy
2. Go through a comparison to another model of the problem other than Random Forest.
3. Engineer at least one feature that is not listed on our initial dataset.

References

This section should include citations for: (1) Any papers on related work mentioned in the introduction. (2) Papers describing methods that you used which were not covered in class. (3) Code or libraries you downloaded and used.

[1] Lin Zhu et al. (2019) *A study on predicting loan default based on the random forest algorithm* Procedia Computer Science, Volume 162, 2019, pp. 503-513. (https://www.sciencedirect.com/science/article/pii/S1877050919320277)