EE 660

Homework 6 (Week 8): Type 1 Project Proposal

Posted: Fri., 10/16/2020 Due: Fri., 10/23/2020, 5:00 PM PDT

This proposal form is for Type 1 projects: Solve a ML problem by implementing a ML system of your own design, using real-world data.

Please fill in both the Project Proposal form (pp. 1-2) and the Dataset Information Form (p. 3). This is required of everyone (each team submits one HW6 with all their names on it). *All fields except "other comments" are required. In each field, replace instructions (black text) with your descriptions. Preferred format is to enter your answers into the Word version of this form, then convert to pdf before submission. If you prefer to use another app instead of Word, then submit a typed version with each field labeled with its title ("Dataset", etc.), and submit as a pdf file.

Please note that this proposal will not be graded like a regular homework. The primary purpose is to give you some feedback on your project topic and plans; the scoring on this homework will be primarily based on whether you put in a reasonable effort and whether the content makes good technical sense.

*Loan Eligibility Using Machine learning

***Project team: Your name(s) and email address(es)**

Weiding Huang
weidingh@usc.edu

***Clear statement of the problem and/or goals.**

Whether a bank can expect clients to pay their credit card on time, the project is aimed to help banks to assess the risk of loans from aspects such as the credit score, annual income, etc. The goal of the project is to evaluate the client in numeric and categorical features and give back if the client can pay off his loan or it is a charged off. Furthermore, we can predict whether an applicant can pay off the loan or not.

***A plan of preprocessing and feature extraction (if applicable)**

We first need to take a look at the feature distribution to make sure the data points are not skewed. After that, the data should be divided into pre_training, training and validation, and test data set. And then we need to normalize all the training data points and test datapoints. Secondly, the data contains a lot of features, many of which might have little to no correlation to the final

decision of the loan. Only important features should be considered. For feature selection, we can try PCA, to try to reduce feature sizes. Also, we can use linear and nonlinear feature selection model such as Pearson's correlation coefficient and Spearman's rank coefficient. There are some features that are not complete, so we should try to fill it using the median or compare with other datapoints that have similar values.

***A plan of your approach**

Firstly, we should find a way to convert all the categorical data to numeric data. (Plan to try panda package) Then, we should use algorithms to determine the weight of each feature. After that, we should try different algorithms such as linear regression, decision tree, random forest, and other algorithms to train the model. Then we can use the trained model to predict whether loans should be approved in the test data set and compare the results. Since the data comes with a test set that does not have the final result, we need to use the test data that is pre-selected from its training data but not yet used. For the original test data, there might be still use such as serving as a base line to the result we have. For the final result, we should expect at least 80% of accuracy since most of the people in the dataset was able to pay off their debt.

***A description of any other work of yours that is related to your class project**

None

***If yours is a team project, roughly describe how work will be divided**

It is a individual project

Other Comments

Anything else you think should be included. For example, if you see potential problems that might arise, you can mention them here with ideas of how you might deal with them. Or, anything that is particular to your project idea that wasn't requested in the form fields above.

Include one form for each dataset you plan to use. (For each dataset's form, you may continue onto an additional page if necessary.)

*Dataset or competition title: Bank Loan status Dataset

*Link: https://www.kaggle.com/zaurbegiev/my-dataset

***Problem type**: regression, classification/logistic regression

***Brief description of dataset and problem domain**: the data set contains

***Number of data points**: 10000

***Number of features or input variables**: 15 features
***Feature or input-variable types**: numeric 10, categorical 5,

***Label (output) type**: binary: whether the client is going to pay off the debt or not

***If Label Type is Categorical, is the number of samples significantly unbalanced (maximal variation of more than a factor of 2)**? Yes,
If yes, *rate as: significant
significant (maximal variation is factor of 2 to 10)
major (factor of 10 to 100)
extreme (more than factor of 100))

***Has Missing Data**? Yes. Some of the annual incomes are missing. If possible, we can use the same annual income of other people who have other similar features

***Is the problem/dataset a Kaggle competition (current or past)?** No

If yes, answer:
***(i) Is the competition current (give the end date), or past**?

***(ii) How much information is available on the Kaggle website (e.g., in "kernels" and links therein)**? Briefly describe what type of information and code is available.

Any other comments on the dataset: