Jeromie Clark

Capstone Proposal

Background

Kaggle is currently hosting a competition on Mortgage Default Risk[1] targeting profiles of individuals who may little to no credit history, or were victims of "untrustworthy lenders". Instead of traditional markers like credit score and payment history, the company uses "alternative data, like telco and transactional information" to determine their customers' creditworthiness.

Problem

Using the dataset provided, which includes both metadata about each borrower as well as the payment history of each loan, successfully create a model that provides an accurate prediction of whether or not a borrower will repay a given loan.

Datasets and Input

The contest sponsor has provided a comprehensive, anonymized dataset that includes any previous credit bureau history and open loan balances, details on any outstanding commercial credit and cash loans, credit card balances, previous applications for loans from the same individual, and a history of all payments made on the loans issued.

Solution

Explore ensemble learning techniques with the goal of producing a model that provides accurate predictions based on our dataset.

Benchmarking

Since this is a supervised learning problem, there is a clear benchmark by which we can compare our performance. In addition, this is an active Kaggle competition, and we can compare the score of this kernel with other proposed kernels in the competition.

Project Design

Data Exploration

- Import the project data and get a feel for what's there by creating labeled tables.
- Learn what fields are typically populated vs. what fields are sparse
- Look at what values fields are populated with and formulate plans for re-encoding and normalizing values as necessary for optimal analysis
- Look at the distribution of values across fields to understand their usefulness as signals
- Calculate some basic statistics about the dataset to answer fundamental questions

[1] https://www.kaggle.com/kailex/tidy-xgb-0-777/data

Data Preparation

- Identify interesting features
- Understand and normalize feature encoding and distribution
- As needed, one-hot encode non-numeric tables

Prepare Training and Testing Data

In this project, it looks like the sponsor has already created training, testing and validation sets; however, I will create them as required using best practices.

Select and Evaluate Candidate Models

Logistic Regression and AdaBoost will be interesting to look at for this project, but I'll probably spend some time to more fully explore the options available in scikit-learn and look for papers in the field that might provide better ideas.

I'll create some prototype models and compare candidates, and select one for further tuning based on my observations.

Model Tuning

I will refer to available literature to identify sensible starting ranges for various hyperparameters in the models used, and employ automated processes like GridSearch for evaluating and tuning those hyperparameters where that makes sense.

Feature Selection

I will calculate the importance of various features in the dataset and select a subset sufficient to produce high-quality results, while optimizing for feasible processing times.

Project Resources

A full description of the Kaggle competition, along with the test dataset can be found at:
https://www.kaggle.com/c/home-credit-default-risk