Udacity Machine Learning Engineer Nanodegree: Capstone Proposal

A proposal for solving an analytical problem by applying machine learning algorithms and techniques.

Matthias Hennig, March 2021

Proposal

The project is a Kaggle competition sponsored by Home Credit Group. Link: https://www.kaggle.com/c/home-credit-default-risk/overview

Domain Background

The project is a typical challenge that financial institutions need to solve: How capable is a borrower to repay a loan granted by a bank? There's a host of research and literature on credit default risk. The Bank of International Settlement in Basel provides a good overview on credit risk on banking: https://www.bis.org/publ/bcbs75.htm

Machine learning techniques can help improve existing models and methods to better assess credit risk.

Problem Statement

Home Credit, like many other banks, provides home loans to people. In order to make decision on who to lend to and at which price (interest rate), Home Credit needs to assess the capability of an applicant to repay the loan. In other words, the bank needs a way to statistically assess the probability that a borrower will not (fully) repay the credit (credit default). This is usually done on checking information and data of the applicant which allow insights into her financial health, employment status, credit history, and any other kind of information that is useful to assess the default risk. The challenge is to find the right data and model to appropriately assess the applicant's creditworthiness.

Datasets and Inputs

Home Credit provides 8 datasets containing information on applicants' financial status and history. The main dataset "application_.csv" contains general information on the customer such as emplyoment status, home ownership, etc and is broken into two files for Train (with TARGET) and Test (without TARGET).

The training data has 307511 observations (each one a separate loan) and 122 features (variables) including the TARGET column. This the label we want to predict that holds a binary variable indicating if a customer repaid a loan (0) or not (1).

Solution Statement

A possible approach would be to train a supervised ML model using provided and newly engineered features from the dataset to see which of these can best predict the TARGET variable. For that we would need to first explore the data to see relations between features and the TARGET variable and afterwards clean, transform, and prepare the data. An additional step might be to reduce the number of features by applying appropriates techniques like Principal Component Analysis. The data is already split into a train and test dataset, so we can train the model, such as an XGBoost classification algorithm, on the former and evaluate its performance on the latter.

Benchmark Model

The results of the model can be benchmarked against actual loan defaults in the past and by continuously backtesting predicted default rates vs the actual loan performance. The datasets already provide an indication of a borrowers past performance, the TARGET column in the test dataset.

Evaluation Metrics

The competition defines the area under the ROC curve (AUC) between the predicted probability and the observed target as evaluation metric. AUC is a measure of how well a model can distinguish between two groups, in this case default and no default. In general, an AUC of 0.5 suggests no discrimination (i.e., chance to correctly predict default is 50:50), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

Project Design

Intended workflow:

1. Data Exploration: retrieve the date, understand the datasets and data types, clean data if necessary (e.g. handle missing values and outliers), check for relations between features, understand distribution of variables

2. Features Transformation: convert variables into features. Standardize/normalize features, apply numerical transformations, perform one-hot encoding, etc.

3. Features Creation: analyse the possibility of deriving new useful features from the existing ones

5. Features Reduction and Selection: select relevant features, reduce main features by extracting principal components

6. Machine Learning Models: apply different strategies, including regression and classification algorithms. For each strategy, optimize with the best choice of algorithm and parameters (try different algorithms, optimize parameters with grid search + cross validation).

7. Evaluation: evaluate the performance of each strategy, and check possibilities of combining them to extract the best of each one and achieving an optimal model.

8. Submit Results