UDACITY CAPSTONE PROPOSAL

A PREPRINT

Sebastian Mack
Machine Learning Nanodegree
Udacity
mack.seb@gmail.com

December 7, 2018

ABSTRACT

Within this document I will propose a Capstone Project which represents the final project in order to graduate in the Udacity Machine Learning Nanodegree program. The idea is to leverage what students learned throughout the Nanodegree program to solve a problem of their choice by applying machine learning algorithms and techniques. This project proposal addresses seven key points (domain background, problem statement, datasets and inputs, solution statement, benchmark model, evaluation metrics, project design) given by the program guidelines.

1 Domain Background

As the field of study I selected an interesting topic from the finance industry that caught my attention when a new competition on the data science platform Kaggle had been announced. The initiator of this contest is one of the largest payment brands in Brazil called Elo (refer to `https://www.cartaoelo.com.br`) which has built partnerships with merchants in order to offer promotions or discounts to cardholders. Main objective of this project is to understand customer loyalty of the anonymized data that Elo is providing.

Typically companies work with databases involving structured datasets which is also the case for our domain. There has been a lot of research and development in order to find algorithms and models for such type of data. In recent years especially a technique called stacking is becoming more popular and is achieving state of the art results within the field. For more information refer to [1, 2, 3, 4]

For me personally this problem is very interesting because it represents a real life problem that not only Elo but many other companies are facing right now. In this situation companies are already acquiring and collecting data with their existing services but still struggle to create a business value out of it. In this particular case we can see that by developing a model that can uncover the signal in customer loyalty, a new business value can be proposed. Both the consumer and the merchants could benefit from a well personalized experience. Furthermore the offering company could gain new clients with this value adding service.

2 Problem Statement

The problem that is to be solved can be described as a supervised machine learning problem because the model will be trained based on a given target feature. Since this target variable has a continuous value it can be further classified as a regression problem. A problem of this type can be solved and modeled with various approaches but in this study the most promising will be applied.

The main problem will be to build a model from the given data that can predict a loyalty score for each card_id given in the test data. Therefor the training data contains several features for each card_id and separate data files with additional information considering transactions of the card as well as information describing the respective merchants of the purchases.

3 Datasets and Inputs

For this project the public available data for the "Elo Merchant Category Recommendation" kaggle competition will be considered. It consists out of the following files:

(`https://www.kaggle.com/c/elo-merchant-category-recommendation/data`)

- **train.csv** - the training set
- **test.csv** - the test set
- **sample_submission.csv** - a sample submission file in the correct format - contains all card_ids you are expected to predict for
- **historical_transactions.csv** - up to 3 months' worth of historical transactions for each card_id
- **merchants.csv** - additional information about all merchants / merchant_ids in the dataset
- **new_merchant_transactions.csv** - two months' worth of data for each card_id containing ALL purchases that card_id made at merchant_ids that were not visited in the historical data.

In addition data field descriptions are provided in **Data_Dictionary.xlsx**.

4 Solution Statement

In order to find a solution to the problem described in the previous section, I will use models and algorithms from supervised machine learning because in our training set we are already given our target variable in form of the loyalty score of each card owner. The specific technique which will be applied to get a model for this project is called stacking. The underlying idea behind stacked generalization was first introduced by a paper [1] from Wolpert. In other research [3] it was shown that super learning approaches provide both a fundamental theoretical as well as practical improvement to the construction of a predictor.

The principals of stacking have not only been proven their potential in acadamic world but also in real applications. In many of the recent kaggle competitions that involved structured data sets the winning models included stacking approaches.

5 Benchmark Model

For benchmarking purposes I decided to apply three different stages in which the main model will be compared to a bench-marking model:

1. Stage: Ensemble model (Gradient Boosting Machine for Regression) vs. stacked model
2. Stage: Stacked model vs. stacked model with tuned parameters
3. Stage: Final stacked model vs. public kaggle leaderboard

6 Evaluation Metrics

Since one objective of this project is also to produce a submission file for the ongoing competition, the choosen evaluation metric will be the Root Mean Squared Error (RMSE) as suggested by the rules.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

In our case we can calculate a score with equation (1) where $\hat{y}$ is the predicted loyalty score for each `card_id`, and y is the actual loyalty score assigned to a `card_id`.

7 Project Design

Finally, I will outline the most important steps within my theorectical workflow in order to find a solution for the described problem. A structured approach will be helpful for a reasonable result and to have a scientific discussion on the final model. My planned workflow includes the following steps:

1. **Exploratory Data Analysis (EDA)**: As a first step I will explore the provided data and make an analysis. This includes summarizing properties and visualizing important outcomes. It will be also very useful to identify features that are relevant for the model and also to give hints for transformations that are required for fitting the data.
2. **Feature Engineering**: Within this step the knowledge gained from the previous step will be applied to clean the data set and to select important features as well as to define new ones.
3. **Train Baseline Model**: When the previous step is completed, the obtained transformed and extended dataset can be used to train the baseline model (or benchmarking model). It will be an implementation of an Ensemble model with gradient boosting for regression.
4. **Train Stacked Model**: The most important step is to train our stacked model which will be combined from several weak learners that still have to be selected.
5. **Tune paramters**: Since there are lots of parameters available in order to train a suffisticated model, it will be necessary to repeat some of the steps and fine tune the model until it is able to produce the desired scores.
6. **Evaluate Metrics**: In the last step, the results and scores of all generated models will be evaluated and compared to one another.

In case I am able to build a model with promising outcomes on the testing data, I will also make a submission file for the kaggle competition to get a score for the public leaderboard to test wether my model can compete with the ones from other kagglers.

References

[1] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[2] Leo Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 07 1996.

[3] Mark Laan, Eric C Polley, and Alan Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6:Article25, 02 2007.

[4] Erin E. LeDell. Scalable ensemble learning and computationally efficient variance estimation. *Doctoral Dissertation*, 2015.