Machine Learning Engineer Nanodegree

Capstone Proposal

Ian Gregson August 2018

Proposal

Domain Background

The domain background for this project is that it comes directly from my work for a startup that is building a customer data platform that helps users maximize their account based recurring revenue. In this field, BizOps profressional leverage large quantities of data from multiple sources such as CRMs or customer engagement analytics tools. The platform we build collates thiss data to provide a single view of an account (a non trivial problem in this space) but much of the insights that are gleaned from this large quantity of data come from the judgements of BizOps professionals studying the data.

The project seeks to take steps into a new way of deriving insights from this large suite of structured data. By automating some aspects of the data analysis with machine learning models will free up the professionals to apply their extensive domain expertise in solving new problems.

I have a personal motivation in undertaking this project because I want to help push my team to become leaders in our field.

Problem Statement

The problem to be solved in this project is the problem of lead, opportunity and account scoring. There is historical data that shows which leads, opportunities and accounts eventually become accounts with annually recurring revenue. BizOps practitioners will use this historical data to make judgements on what active leads, oppoortunities and accounts are most likely to convert by comparing their features with the features of examples from the historical data.

This project seeks to build a machine learning model that will make these judgements.

Datasets and Inputs

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

The datasets that are to be used in this project come from several customers of my company. Each

customer has a number of leads, opportunities and accounts. For each of the categories, each record has a binary classification state: converted or did not convert. This historical data provides a labelled dataset for each category that will allow us to build a model that can "score" new data i.e. assess with what probability each record belongs to the "converted" classification.

These datasets are appropriate to the problem because they are real data collected from companies who are actively engaged in sales processes. The data has been collected in and generated by various software systems that the buesinesses use to support their sales processes. This data has then been collated in the Rekener (my company) customer data platform.

The shape of each dataset will be examined early in the project, but my initial intuition is that the data will be very wide (i.e. will have many features) and will be sparse.

Solution Statement

The solution to the problem described above is to build a binary classification model that will predict with what probability a lead, opportunity account will or will not convert. The model's performance can then be measure with an F1 score metric and compared with the performance of a dummy classifier for reference.

Benchmark Model

The benchmark model for this problem will be a DummyClassifier. Since there is no existing model for predicting conversion rates in the Rekener data, the DummyClassifier will provide an adequate baseline and help establish whether or not the solution model is better than a guess.

Though this sounds sub-optimal, this is actually reflective of the real world where BizOps practitioners will make an informed guess about whether ot not given leads and opportunities will convert by consuming the given data and applying their intuition gained from years of experience working in the domain.

This benchmark model will not allow for any assertion to be made on whether or not the solution model is an improvement on the BizOps professional intuition but, as outlined in the solution statement, the purpose of this project is to create a model that is known to be reliable enough for the BizOps professional to consult with in order to augment their experience and intuition and help them converge on their predictions more quickly. A "better than a guess" benchmark is very useful in establishing this criteria.

Evaluation Metrics

Since this project aims to solve a classification problem, I expect that Confusion Matrix and F1 Score metric will provide suitable evaluation of the performance of the benchmark and solution models.

Project Design

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the

data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

A theoretical workflow for this project can be summarized as follows:

Basic analysis of the shape and features of the data

I expect the data to be in a format that will require a great deal of preprocessing. In this first step, I will asses the shape of the data and the features it contains and identify features that need to scaled, one-hot-encoded, fillna or removed.

Once these steps are identified, I will apply all the necessary data transformations.

Preparation of evaluation metrics

In this step I will import the necessary libraries and initliaze the F1 Score and Confusion Matrix metrics.

Perparation of dummy classifier

Next, I will prepare the dummy classifier that will be used as the baseline for evaluating model performance.

Initialization of models and hyperparameter search

Next, I will initialize the models that will be used in this project. Since I will be working with structured data, I expect that there will be a high number of features and the datasets will have the potential to be sparsely populated. From my study on this course, I have gained the intuition that the most appropriate models will therefore be Support Vector Machines and Random Forest.

I will use Grid Search to locate the best hyperparameter configuration for both models.

Training

In this section I will prepare training and validation datasets, then train each model. I will save each model as a pickle for future inference.

Evaluation

In this section I will asses which model performed best based on the F1 Score metric. The model with best performance can then be compared with the performance of the dummy classifier to establish whether or not the model provides a better than random classification.