Machine Learning Engineer Nanodegree Program

Capstone Proposal

Domain background: A key aspect in the marketing strategy of any company is to understand the value and the needs of its customers. Determine which customers the company should focus and how much to spend with promotions, marketing events, campaigns and other related activities is a fundamental factor to its success.

Problem statement: In this context, our problem is to (1) understand what customers are more valuable to the business and (2) how much the company should spend with marketing and promotions for each customer.

Datasets and inputs: The data for this project is the basket-level transaction history for each individual customer. This dataset is from an old kaggle competition and has three relational tables:

> **transactions.csv** - transaction history for all customers
>
> **offers.csv** - incentive offered to each customer and information about the behavioral response to the offer
>
> **offers_information.csv** - contains information about the offers

Note: The data and the description for each field can be accessed through the link above. Considering the competition rules, I am not providing a sample of the data.

Solution statement: To answer both questions, we can estimate the customer lifetime value (CLTV). This metric is given by the customer revenue (total purchases) minus all the costs related to acquiring and serving the customer. In this project, there is no information about costs and therefore the revenue for the next 3 months will be considered as our CLTV target.

The solution to this problem will be a machine learning model that can predict customer revenue for the next 3 months using the individual transaction history.

Ideally, the prediction window would be larger, but we are limited to a one year transaction history.

Benchmark model: The machine learning solution will be evaluated in an out-of-time sample. It will also be compared with the classic approach for CLTV modeling which consists in fitting one or more parametric distributions to the transaction history. More specifically, the machine learning results will be compared with the Pareto/GGG model implemented in the Lifetimes python package.

Evaluation metrics: The predicted revenue will be compared with the observed value for both the training set and test set (out-of-time sample). The revenue difference between customers can be quite large, therefore I will be using **Mean squared log error (MSLE)** or **Mean absolute percentage error (MAPE)** to evaluate the results.

Project design: The final product will be a deployed machine learning model that can be used to predict the customer revenue for the next 90 days. It will possibly also include a dashboard to display the customer transaction history analysis and predictions given its id. The project will follow these general steps:

- Upload raw data to S3.
- Initial data exploration using a random sample of customers.
- Data cleaning and preprocessing.
- Feature engineering and target extraction using PySparkProcessor to build an ETL with the full dataset (iterative process since we can build more features).
- Baseline model and analysis (GLM and/or Pareto/GGG model).
- Test different machine learning algorithms.
- Build final model and hyperparameter tunning.
- Evaluate model score and features (performance and stability).
- Deploy the model
- (optional) Build dashboard
- Documentation