Machine Learning Engineer Nanodegree

Capstone Proposal

Jesus Martinez

July 20th, 2018

Proposal

Domain Background

“La Petite” Jewelry opened its doors on the summer of 1979 in the city of Monterrey, Mexico. The founders saw an opportunity as they had a close relationship with Mexico’s most important jewelry manufacturers and owned well located commercial premises. They currently offer a great variety of products (both in silver and gold) ranging from rings, bracelets, neck chains (most sold product according to owner), earrings, etc.

Keeping track in how over a 1,000 different products are being sold on a daily basis can be a daunting task, especially if technology based solutions have not being widely adopted. As it can be clear, the problems derived from a deficient monitoring of products being sold can be countless therefore we will only focus in trying to answer a simple question. **What quantities should the owner order from an specific product for maximizing utility purposes?**

Problem Statement

The question previously defined is asked every month by the owner and using personal intuition and advice from in-house sales representatives they come up with an amount of product to be ordered. As expected this process may be quite practical however not so accurate as in average it deviates between 65% to 75% of exceeding product always. Generating a more formal and accurate solution without affecting practicality will be the main objective.

Datasets and Inputs

Our dataset to be used consists of 50 data points representing a particular day (starting in April 10 to May 29, 2018) and the amount sold for a single product.

The data was obtained from a temporary installed Point of Sale system (as this was never used before) serving for two main purposes one of them being, to calculate the current strategy error and also it will be our main artifact to be used when generating new predictions.

Solution Statement

The solution proposed aims to provide on a monthly basis a prediction for the amount to be ordered for the product in concern. It will provide a level or percentage of accuracy (against standard and simple metrics) for the user to quantify the risk involved.

Benchmark Model

Existing methods within the business reside in coming up with a number that avoids falling in a shortage of the product using intuition, this accuracy ranges from 65% to 75% in average. It's worth to mention that no data was retrieved properly in the past on a daily basis. From this perspective our goal is to come up with a model with at least 60% accuracy.

Evaluation Metrics

Standard accuracy metrics will be taken into account for this project such as R^2 (coefficient of determination) and MAE (mean absolute value) percentage error. The former will aid us determining the best model, from those to be considered only, and its comparison with calculating the mean of all target variable values. Once the previous step is defined MAE percentage error is useful when understanding how far are we from the true value.

Project Design

Our main goal in this project is to predict product purchases for posterior months in order to make informed decisions when acquiring these products from jewelry manufacturers. Current owner procurement decisions exceed between 65% to 75% of what is actually sold in which it may seem a clear case of a low hanging fruit scenario. With this in mind we will aim to at least obtain a 60% accuracy.

The first step to consider consists of cleaning out the data removing, if any, values that do not align or make sense and proceeding with ***obtaining general descriptive statistics*** such as minimum, maximum, mean, median and standard deviation from the "grams" column or target variable. Once this is completed a general visualization chart will be created plotting the variable named earlier vs the dates in an effort to find any observable patterns.

The next phase refers to finding relevant and reproducible features from within our dataset named formally as ***feature extraction***. Principle characteristics that would be searched relate to a balance between high representability and high frequency these can be such as days of the week (Monday, Tuesday…), weekdays vs weekends, etc. Examples of features that will not be considered as they do not fall under the previous description can be holidays, paydays, specific weather conditions like rainy days, etc. This decision resides in the fact that with a low volume of data, considering features with low frequency may not be picked up by our models making it a futile activity.

After choosing our best set of features we will continue in ***modeling and evaluating*** them separately against two machine learning algorithms (linear regression and gradient boosting regressor) plus a simple prediction using the mean of the target variable values. Similar to our previous conclusion within our feature extraction phase, considering these variables jointly is expected to bring less value than independently. To summarize this approach, in a case where "weekday" and "weekday-end" is selected a total of five models will be processed. Using a linear regression and gradient boosting with "weekday" versus "product grams" and "weekday-end" versus "product grams" respectively. As explained on the Evaluation Metrics section the best model to continue further parameter tuning will be defined by our metrics R^2 and MAE.

A document containing the ***conclusions and next steps*** from this analysis will be delivered to the customer and a new proposal(s) could be discussed if required.