Product Cost Predictor

Capstone Project Proposal

DOMAIN BACKGROUND

Student briefly details background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited. A discussion of the student's personal motivation for investigating a particular problem in the domain is encouraged but not required.

The USMCA (Canada, Mexico, US) industrial fastener market is significant. It is valued at $19B today and is expected to grow at a 3% CAGR reaching $22B by 2025 [1] . It's a competitive market too with competent manufacturers and demanding customers that expect quick, competitive cost proposals for their custom requirements. Today, custom fastener products require subject matter experts to accurately estimate the cost of each customer request. This takes time with current tools and process.

As a product manager for a major fastener manufacturer, I am caught between timeliness and accuracy of cost proposals. On the one hand, a timely cost estimate will please a customer. On the other hand, an inaccurate cost estimate can frustrate a customer or backfire on the fastener manufacturer if the cost is too low! A method to provide timely and accurate economic proposals is needed if our company wants to grow with this competitive market.

PROBLEM STATEMENT

Student clearly describes the problem that is to be solved. The problem is well defined and has at least one relevant potential solution. Additionally, the problem is quantifiable, measurable, and replicable.

It takes a week from the receipt of a customer request to respond with an economic proposal. Most of the time is used researching each feature cost, summing the individual costs and validating the total cost. Although it is accurate, the time to deliver the economic proposal exceeds the three-day preference of most major customers. Delivering an accurate economic proposal within three days would open more opportunities to win new business. One solution is to apply machine model regression to predict a cost using the features and costs of similar parts we make today.

DATA SET & INPUTS

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. 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.

Details of currently shipped product are continuously updated to an excel file that documents the profitability of each plant within my company. This file could be repurposed as a training and testing set for my cost predictor! The benefit of this file is all the features required to calculate cost are defined for each record along with the actual cost. The downside is the file has 460 columns and 55,000 rows. Many cells are empty, and the data can be inconsistent.

SOLUTION STATEMENT

Student clearly describes a solution to the problem. The solution is applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, the solution is quantifiable, measurable, and replicable.

A four-step workflow will be used to achieve a solution. The FIRST step is to refine the input data. Refining techniques will include examining histograms of numerical values to identify and remove outliers. NANS will be eliminated to create a fully populated data set. Categorical data will be OHE and numerical data will be scaled to eliminate any issues that could arise with the order of magnitude for the different values. The last step will be to separate the data 80% to training and 20% to testing.

The SECOND step will be to instantiate, train and test machine learning algorithms. I will experiment with KNN, Decision Tree, Random Forest, AdaBoost and Linear Regression and Pytorch. The most accurate candidate will go onto the THIRD step to optimize hyper parameters to improve the results as much as possible. The FOURTH step will be to create an AWS endpoint where the preferred model can be accessed.

BENCHMARK MODEL

A benchmark model is provided that relates to the domain, problem statement, and intended solution. Ideally, the student's benchmark model provides context for existing methods or known information in the domain and problem given, which can then be objectively compared to the student's solution. The benchmark model is clearly defined and measurable.

A Decision Tree Regressor with default hyperparameters will be used as the benchmark. Its accuracy will be the baseline other models will be compared against. A perfect model would exactly predict the cost defined for each product on each row based on the column features. There are thirteen feature columns out of the four hundred-sixty profitability columns. The categorical features are:

FEATURE	DESCRIPTION	COUNT
CLASS	product type: bolt, nut, washer, ..	16
SUB	further definition of the class	64
ASSY	An assembly description if it is made from more than one part	7
HEAD	head configuration : hex, socket, ..	12
DRIVE	Drive style : Philips, Torx, ..	16
THREAD	Thread geometry: machine, trilobe, PT,..	28
NOM	Diameter :M6 , M10, 1/2”, …	31
POINT	Point configuration: M-Point, dogpoint,..	21
HEAT	Heat Treat & tempering	14
LOCK	Torque Retention, Nylon, Crimp,..	10
PLATE	Surface Treatment: Zn, ZnNi ,..	216

These describe the physical characteristics of each part so that someone skilled in fastener production would know what is required to make each part. There are several to many variations within each feature that make endless combinations although there are common patterns. Two continuous features “qty”, the annual number of required pieces and “mm”, part length are the integer and float values respectively.

EVALUATION METRIC

Student proposes at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model presented. The evaluation metric(s) proposed are appropriate given the context of the data, the problem statement, and the intended solution.

The evaluation metric is the Mean Square Error (MSE), Mean Absolute Error (MAE), R2 Score. Although I would like to set 90% as the target, I will have to dig into the data to see what is possible. I believe that some product will be more accurate than others because there will be more instances to train upon.

PROJECT DESIGN

Student summarizes a theoretical workflow for approaching a solution given the problem. Discussion is made as to what strategies may be employed, what analysis of the data might be required, or which algorithms will be considered. The workflow and discussion provided align with the qualities of the project. Small visualizations, pseudocode, or diagrams are encouraged but not required

The exceptionally large profit data spreadsheet must be culled down to the appropriate feature data. That feature data must be anonymized to maintain the intellectual property while providing enough granularity for review by Udacity

The baseline input data will be wrangled into a set of critical features and a response cost. Records with missing data will be eliminated. The categorical data will be one-hot-encoded and then Primary Component Analysis will be applied to identify primary features. A cross variance matrix will be plotted to identify components that are most orthogonal and will span the solution space most effectively. The final, clean data will be used optimized in in three different regression models. The most accurate model will be chosen as the final preferred predictor.

PRESENTATION

Proposal follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used and referenced are properly cited.

The Workflow will follow this approach:

1. LOAD & EXPLORE DATA
   a. Segment features and response
   b. Outlier , NAN Removal
   c. OHE categorical data
   d. Train , Test Split 80% , 20%
2. MODEL APPLICATION
   a. Standard Scaler
   b. INSTANTIATE: KNN, Decision Tree, Random Forest, AdaBoost and Linear Regression and Pytorch.
   c. Train models
   d. Test models
   e. Compare Accuracy
3. Optimize the hyperparameters of the best candidate model
   a. Grid Search CVBatch, epoch, learning rate , hidden size if Pytorch
4. Create AWS of final model

References

[1] Freedonia, "Global Industrial Fasteners," Freedonia, Cleveland, 2017.