Sales Forecasting Using High-Performance Computing

Sowmya Bhatraju, Pavan Ghantasala, Yu Lin Tai, Ashutosh Porwal, Swati Srivastava, Vineet Suhas Soni, Yang Wang

Purdue University, Krannert School of Management

sbhatraj@purdue.edu; nghantas@purdue.edu; tai21@purdue.edu; porwal@purdue.edu; srivas98@purdue.edu; vsoni@purdue.edu; yangwang@purdue.edu

ABSTRACT

The research is intended to improve inventory management and assortment planning of a national auto-parts dealer through reliable and efficient sales forecasting. Exploratory data analysis was performed on three categories of data- batteries, brakes, and filters, for the past two years at a unique SKU-Store combination. High-performance computing was performed using Bell Cluster and Dask machine learning libraries. Several feature engineering experiments, PCA analyses, etc. were carried out to build robust models such as linear regression, lasso regression, and OLS. The final regression model selected has the highest adjusted R-square and least mean square error, which were the two metrics defined for model selection. This model had the highest interpretability and lowest run time.

INTRODUCTION

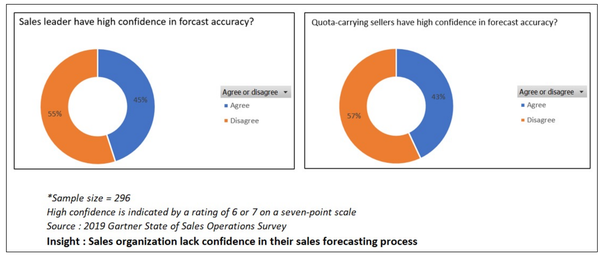

Fig 1. % of correspondents indicating if there is a high confidence in forecast accuracy

• What is the way to forecast store sales using different regression techniques including interaction terms and feature engineering ?
• Can you show how using HPC capabilities would allow one to eventually find an optimal robust linear model that performs much better than trying a generic backward selection type approach?

LITERATURE REVIEW

Sales forecasting is the process of predicting the number of sales a firm will generate within a specific timeframe such as a week, month, quarter, or year. Creating a predictive model to forecast sales with high dimensionality issues requires careful consideration so that model computation time is within an acceptable time frame. Multiple variable selection methods, in cases of many explanatory variables, can be combined, and the union of selected variables from those methods can be taken to reduce the risk of eliminating variables having a significant impact on the response variable. PCA can be performed to deal with high dimensionality, but it runs the risk of eliminating some features that might be strong predictors of the dependent variable. A way around this is to categorize variables in broad groups according to their common attributes and run PCA on each group separately to reduce the likelihood of missing out on important predictor variables. Linear Regression has been the optimal choice for sales prediction for practitioners finding the balance between interpretability, time complexity, and accuracy. It is easy, simple, and feasible to fit at SKU level high dimensional space, with approaches to non-linearities such as multiplicative interactions log and exponential transformations applied to variable space.

METHODOLOGY

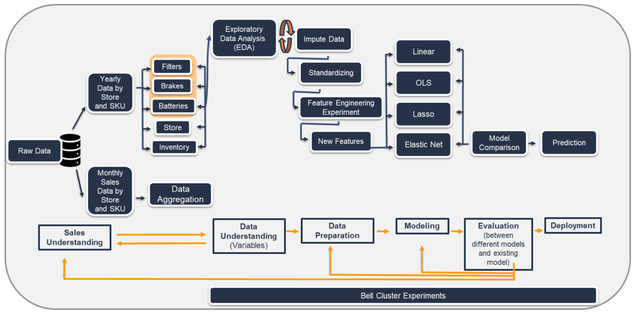

Fig 2. Study Design

EXPLORATORY DATA ANALYSIS

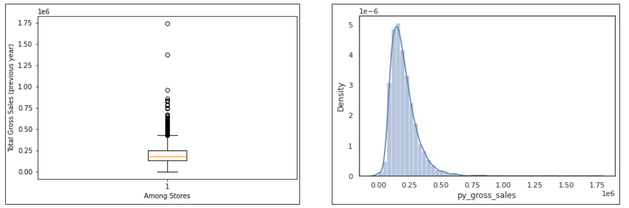

Fig 3. Distribution of the previous year total gross sales at store level

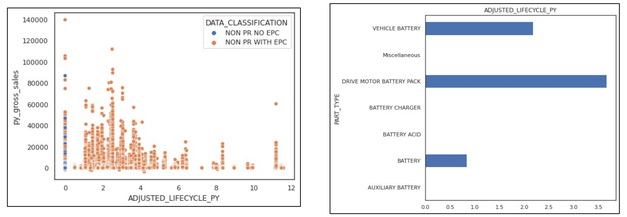

Fig 4. Relationship b/w lifecycle and previous year gross sales for different parts

The bar chart shows the lifecycle of different part types, with "drive motor battery pack" having the highest average adjusted lifecycle. From the scatter plot, we could tell that parts classified as “Non PR No EPC” had relatively lower sales compared to part with EPC.

STATISTICAL RESULTS

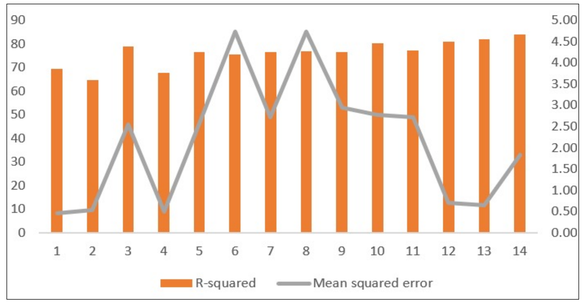

Fig 5. Model results

The graph illustrates the best few experimental models with high adjusted R-square and low mean squared error. The adjusted R-square ranges from 64.6% to 83.8% and MSE was between 0.465 and 4.722.

- Models such as Linear regression(1,2,12), Lasso(4) and OLS(3,5-11,13-14) were explored. Interaction terms were created by combining demographic variables, lost sales and quantities in the previous year, quantity sold in the previous and previous-to-previous year. PCA analysis was also performed.
- Model 1(Linear regression model) had the minimum MSE of 0.47 but the predictors used in the model were not significant which led to a lesser R-square value. A few OLS models (6,8) had a decent R-square of ~75% but had a high MSE of 4.72 indicating a high biased or high variance estimate. The best model obtained was an OLS model (13) which used the interaction between quantity sold in previous and previous-to-previous year, with an adjusted R-square of 83.8% and Mean Absolute Error of 0.14.

EXPECTED BUSINESS IMPACT

Through our use of high-performance computing on bell cluster, we were able to perform hundreds of experiments for different models and reached an MAE of <1 with a fairly good adjusted R-square value of >80%. Our model could play a crucial role in creating a strong business impact for the company. Below are the monetary benefits that could be availed through our model:

- Through percentage decrease of mean absolute error by ~90%, sales could be more accurately predicted. This would further reduce the inventory management costs for the company.
- With model-run time being as low as 2 min, it can provide another cost-saving opportunity

ACKNOWLEDGEMENTS

We would like to thank the Data Science Team of the company we worked with and Professor Yang Wang for her guidance and support on this project.