To Stock or Not to Stock: Forecasting Demand in Grocery Stores

Zhenghao Ye, Yixuan Shi, Andrew Lee, Yidan Gao, Matthew A. Lanham
Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907
ye122@purdue.edu, shi280@purdue.edu, lee2061@purdue.edu, gao311@purdue.edu, lanhamm@purdue.edu

Abstract

This study uses descriptive and predictive data analysis methods to create a useable solution to solve the problem faced by retail companies, which is how to accurately forecast demand and determine the right time, right amount, and right target demographic to sell products. First, we filtered and processed our observed data and used a GA (genetic algorithm) to test if there is a significant influence from perishable goods. They stated that as of now most of the current methodologies for computing sales forecasts of short shelf life products rely on linear or nonlinear time series algorithms. With our revised data, we utilize models such as neural networks, GLM, linear regression, and others to address the problem.

Introduction

Knowing when and what to sell to the customer at the right time and right location is essential to achieve maximum sales. Successful profit optimization entail both cutting down costs and understanding when to surge in sales volume. While cost control can be achieved by variable and fixed cost reduction, in the premise that products are under plausible quality, businesses should also grip viable sales volume increases by utilizing both experimental acumen and technological optimization support.

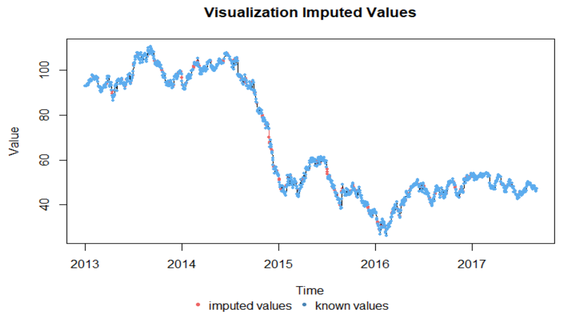


Figure 1. Oil Prices by Time (w/ Imputed Values)

Literature Review

Study	GA	Regression Tree	Neural Network	Clustering & Hierarchical Forecasting	XGBoost
Doganiz, Alexandridis, Patrinos, & Sarimveis, 2006	✓				
Ali, Sayin Woensel, & Fransoo, 2009		✓			
Zhang, 2003			✓		
Heober, Gossmann, & Stuckenschmidt, 2017			✓	✓	
Jason Brownlee, 2016					✓
Our Study	✓	✓	✓	✓	✓

We found that this topic has been attempted to be solved throughout the years and each article utilized a different method to come up with their results. Most commonly, we found that the use of ARIMA, ANNs, and Tree models were used.

In our solution, we use many of the same models used in each of these studies, while also adding in our own choice of models that we believe will create a nice solution to the problem.

Table 1. Literature review summary by method used

Methodology

The figure below summaries our process for our study. First, we pre-processed the sampled data from original data, and partitioned it into train and test sets for cross-validation. Before actually modeling the data, we did principal components analysis using different approaches in order to validate the features we are using. Then the data is transferred to different models. After the modeling we did k-folds cross-validation and created plots of the results to evaluate each models accuracy.

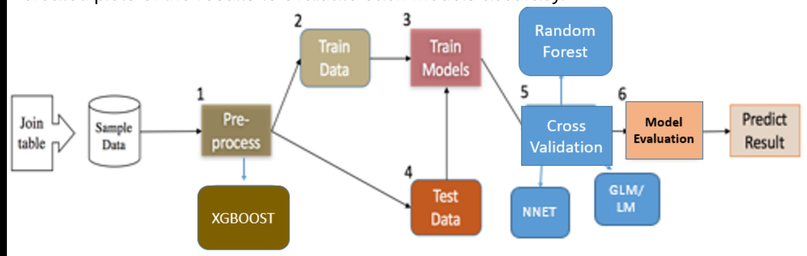


Figure 2. Process Design

Data Cleaning & Pre-Processing

We imputed missing values in oil price with the R **mice** package because oil prices tend to form a flatter pattern over time. We also made sure that the missing values in oil price was not too large that it would make the data set skewed. In regards to other missing values in our sample dataset, we deleted them because they did not reflect a large portion of our data.

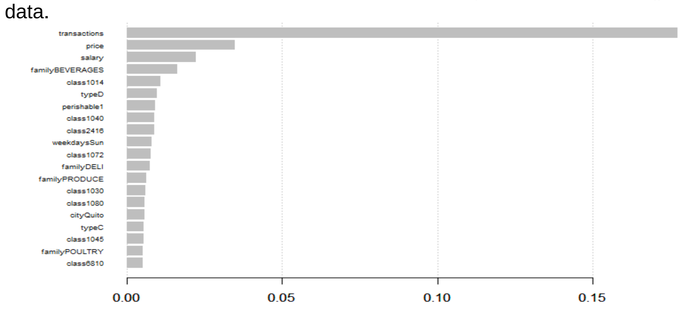


Figure 3. Top 20 Selected Variables

Feature Selection

Different from the traditional way of excluding near-zero and highly correlated variables, we used XGBoost instead to select what features should be included in the model. We put all of our features into XGBoost, in order to select the top 20 features that should be included in designing our models. XGBoost builds the model in stage-wise fashion in order to prioritize which features are the most important.

Model Design

In designing our project, we decided to partition the data into 80-20% train-test sets, and used 5-fold cross-validation. We decided to partition the data in this way as it gives enough of values to train and obtain a robust model, while providing the ability to estimate the test error more accurately.

Methodology (Approach) Selection

Linear Regression
Neural Network
Generalized Linear Model
XGBoost

Results

Model Evaluation / Statistical & Business Performance Measures

The predictive models were evaluated on the RMSE statistical performance measure because it was the main performance measure we discussed in class to use when dealing with regression type problems. When given our case, the case mentioned in its objective statement that the best model would be based on a Normalized Weighted Root Mean Squared Logarithmic Error, so we used RMSE as a more simplified version of what was desired. This metric is suitable when predicting values across a large range. RMSE will give high weights to large errors, and we want this as large errors are undesirable in our business scenario.

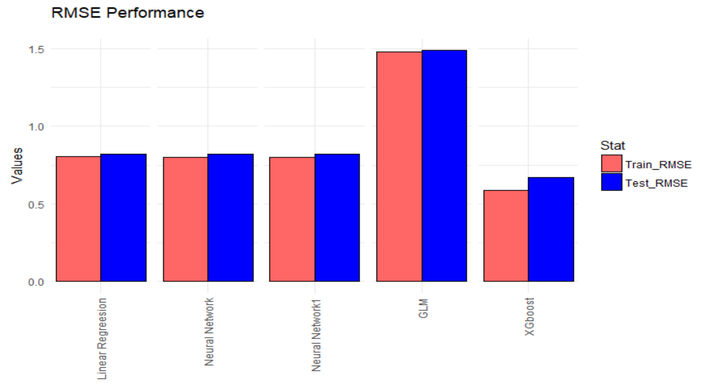
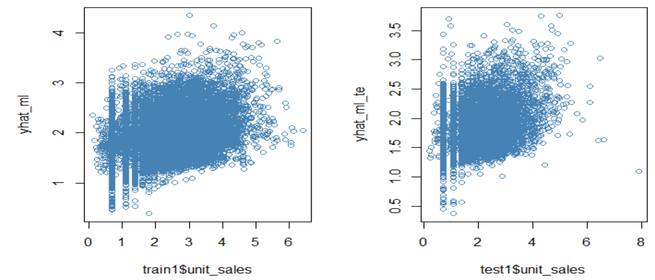


Figure 4. Model Evaluation

Figure 5. Predicted vs Actual for Linear Regression



Conclusions

Based on the results given by our five models, we can conclude that the best model when forecasting demand, from our research was our linear regression model.

Transactions	$2.539e^{-04}$
Price	$-1.080e^{-03}$
Salary	$-4.843e^{-03}$
Type D	$4.006e^{-02}$
Perishable	$2.631e^{-01}$
Weekdays (Sunday)	$1.564e^{-01}$

We can conclude that from our research that there exists a linear relationship between unit sales and parameters such as transactions, family beverages, salary, etc.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.