IE4211 Modeling and Analytics
AY2019/2020 Semester 2

Group Project

Group 24:

Cherie Sukhita Irawan (A0159715W)
Ng Jing Hui Darrell (A0155350M)
Tan Ci Kang (A0159931W)

Table of Contents

1 Data Exploration 2

1.1 Data Processing 2

1.2 Methods to Predict Sales 2

1.2.1 Multiple Linear Regression 3

1.2.2 Subset Selection 3

1.2.3 Shrinkage Methods 3

1.2.4 Tree-Based Models 4

1.2.5 Support Vector Machines 4

1.2.6 Summary of Results 5

1.3 Insights 5

2 Inventory Decision 6

2.1 Predicted Sales and Demand 6

2.2 Optimal Inventory 6

3 Summary 7

1 Data Exploration

1.1 Data Processing

After looking into the data, it was found that each ‘brandID’ has multiple ‘productID’, but each ‘productID’ has the same ‘brandID’. Therefore, ‘brandID’ is removed from the dataset since ‘productID’ can be used as identification of all the products. This means that the information on ‘productID’ is useful whereas the information on ‘brandID’ is not as useful.

Since there are various features in the data provided, it is helpful and necessary to treat some features as categorical variables. Scatter plot and box plot are used to visualise the relationship between the features as shown in the Jupyter Notebook.

The variables, ‘productID’, ‘weekday’, ‘attribute1’ and ‘attribute2’, are identified as categorical variables. The variable ‘productID’ is used to identify the various products, whereas the variable ‘weekday’ is used to identify the day of the week. They are therefore considered as categorical variables. Additionally, the variables ‘attribute1’ and ‘attribute2’ only take a few possible discrete values and will be taken as categorical variables in this project.

To further investigate the relationship between the attributes and the predictor, the correlation coefficients are also found. Through the scatter plot and the correlation coefficients, we found that there is significant relationship between ‘attribute1’ and ‘attribute2’, ‘ma14SalesVolume’ and ‘clickVolume’, as well as ‘avgFinalUnitPrice’ and ‘avgOriginalUnitPrice’. The variables ‘avgOriginalUnitPrice’ and ‘avgFinalUnitPrice’ are highly correlated with a value of 0.899. This is because ‘avgFinalUnitPrice’ is derived from applying discounts and coupons on ‘avgOriginalUnitPrice’. This means that ‘avgOriginalUnitPrice’ and ‘avgFinalUnitPrice’ are similar to each other. Furthermore, since users only look at the average final unit price when making a decision, ‘avgOriginalUnitPrice’ is removed from the dataset. For ‘ma14SalesVolume’ and ‘clickVolume’, it makes sense that they have high correlation because higher number of people that click on the product would mean that more people are interested in the product, which results in more sales. More clicks over the last 14 days would lead to more sales in the last 14 days.

1.2 Methods to Predict Sales

1.2.1 Multiple Linear Regression

Linear regression is one of the ways to model a relationship between the response and the predictors. Due to the simplicity of the model, linear regression is used as the first method for most analytics projects. This is because a linear model usually predicts well enough and provides better insights than other models which are more complicated. Hence, multiple linear regression is performed. The ordinary least squares model produces an R-squared value of 0.713 and adjusted R-squared value of 0.702. This indicates that linear regression is a relatively good model. However, there are many variables which have a p-value more than 0.05. A p-value of less than 0.05 indicates that the predictor is statistically significant, whereas a p-value of more than 0.05 indicates that the predictor is statistically insignificant for predicting sales. In other words, variables which have a p-value of more than 0.05 means that they are unlikely to have a linear relationship with sales.

1.2.2 Subset Selection

Despite the advantages of linear regression, some issues are associated with it. In some cases, some of the predictors are not associated with the sales. Therefore, subset selection is very important for a model with a large number of predictors as it selects the important features which are the most significant in the model. By including the important features, it will help to improve the accuracy of the predictions. Using the RSE package on Python, it is found that the best model with the lowest mean squared error (MSE) uses all the features. The lowest MSE for this model is 1375.284 and an accuracy rate of 0.734. Subset selection indicates that all the features are statistically significant. This is not surprising considering that this is a low-dimensional model.

1.2.3 Shrinkage Methods

Furthermore, the dimension of β may also be too large and therefore we can use regularization to reduce the search space. Other than improving the accuracy of the prediction, the interpretability of the model can also be improved. To reduce model complexity, LASSO and Ridge Regression are applied to shrink the coefficients. Ridge Regression is able to shrink the coefficients while LASSO Regression is able to do feature selection by forcing some coefficient estimates to be exactly 0. To apply this method, the data has to be standardised as the predictors are in different units. Cross validation is also conducted to choose the best parameter alpha. The best LASSO parameter alpha is 1 and it has an MSE of 1314.610. The best Ridge parameter alpha is 100 and it has an MSE of 1395.856. The results found that for this project, LASSO works better than Linear Regression and Ridge Regression.

1.2.4 Tree-Based Models

Decision tree methods allow for the stratification of predictor space into a few simple regions. This method is accompanied with Bagging, Random Forest and Boosting. For tree-based models, the data is not standardised. Cross validation is also performed to find the best cost complexity pruning (CCP) alpha. The best CCP alpha for Decision Tree is 10, with an MSE of 2479.527.

With Bagging, the best ccp alpha is 0.01, with an MSE of 999.205. For Random Forest Regressor, the best ccp alpha is 0.01 and it has an MSE of 999.155. Lastly, for Boosting, cross validation is conducted to find the best learning rate parameter. It is found that the best learning rate parameter is 0.07 and it has an MSE of 1367.886. Hence, the best method for tree-based models with the lowest MSE is Random Forest Regressor.

1.2.5 Support Vector Machines

Support Vector Machine is applied for regression problems as Support Vector Regression (SVR). It separates the features into various regions by a hyperplane. It also transforms the feature space using kernels. There are three different kernels used in the project, ‘linear’, ‘polynomial’ and ‘radial basis function’. Cross validation is also performed to find the best C, which is the regularization parameter. The data is also standardized for this model.

For the ‘linear’ kernel, the best C parameter is 1 with an MSE of 1087.751. The best C parameter for ‘polynomial’ kernel with degree 2 is 100 with an MSE of 1156.618. At last, for the ‘radial basis function’ kernel, the best C parameter is 100 with an MSE of 1183.883. For the ‘polynomial’ and ‘radial’ kernel, the highest C parameter gives the best model. This means that it is not appropriate to transform the feature space and the features have a better linear relationship to the response compared to polynomial or exponential.

1.2.6 Summary of Results

Method	Alpha	CCP Alpha	Learning Rate	C	Mean Squared Error (MSE)
Multiple Linear Regression					1375.284
LASSO	1				1314.610
Ridge Regression	100				1395.856
Decision Tree		10			2479.527
Bagging		0.01			999.205
Random Forest Regressor		0.01			999.155
Boosting			0.07		1367.886
Support Vector Machine (Linear)				1	1087.751
Support Vector Machine (Polynomial with Degree 2)				100	1156.618
Support Vector Machine (Radial)				100	1183.883

1.3 Insights

From the Principal Component Analysis (PCA) biplots, several insights can be deduced.

The features 'meanPurchasePower' and ‘meanCityLevel’ have a strong correlation. Customers from a higher city level tend to have higher purchasing power. This makes sense because larger cities would mean that it is more developed and would have a stronger economy, thus its population would have a higher purchasing power. They can increase their revenue by expanding into other areas with a similar city level of their products.

As mentioned earlier, ‘clickVolume’ and ‘ma14SalesVolume’ have a strong correlation. When there is a higher click volume, more people are interested in the product, resulting in higher sales volume.

There is a strong correlation between ‘meanEducation’, ‘plus’, ‘MaritalStatus’, ‘meanAge’ and ‘meanUserLevel’. There is a lot of information about the target customers that can be deduced from these features. This information is useful in building a customer profile for the company. Members of the company who are also customers are more likely to have a mean higher educational level. This makes sense that the company employees are more educated than the general population of customers. Customers who are members of the company have a higher user level. This makes sense for the company employees to believe in the product of their company. It could also be due to employee discounts so that employees choose to purchase products from their company instead of similar products from other companies. For customers with a higher user level, they are more likely to be married and older. There is also a strong correlation between ‘meanEducation’ and ‘meanUserLevel’. Therefore, the company could increase their sales by marketing towards families and working adults. Their target demographic to generate the most sales would be the middle income working adults in developed cities.

2 Inventory Decision

2.1 Predicted Sales and Demand

Since Random Forest Regressor with CCP alpha of 0.01 has the lowest MSE, this model is chosen to predict the sales of the products in the test dataset. A histogram is also plotted to visualise the predicted demand of the test dataset. From the histogram, it is concluded that predicted demand is exponentially distributed. This is further validated by comparing it to the distribution of the true demand in the training dataset.

2.2 Optimal Inventory

Given that the product cost is $12 and the salvage price is $8, the overage cost is $4. Since the price of the product is $20, the underage cost is $8. The inventory decision for each data point is then calculated based on the exponential distribution of predicted demand using

$$\text{optimal inventory} = \frac{-ln(1-p)}{\lambda} \text{ where } p = \frac{C_u}{C_u+C_o}$$

This value is rounded up in order to meet the demand. The inventory model uses the same model as the prediction model.

3 Summary

In conclusion, patterns from the Data-train.csv are first analysed using scatter plot and box plots to find the relationships between the features. Some of the features are also omitted to reduce the model complexity and some of the features are converted into categorical variables to improve model accuracy. Multiple regression techniques are also performed to predict the sales for the Data-train.csv. The various models used to predict the sales are multiple linear regression, LASSO, Ridge Regression, Decision Tree, Bagging, Random Forest Regressor, Boosting and Support Vector Machines. Cross validation is performed for each model to find the best parameter with the lowest MSE. Since a Random Forest Regressor with CCP alpha of 0.01 has the lowest mean squared error, it is chosen as the best method to predict the sales for the given test dataset.

From PCA, we can study the relationship between the features. The biplot of the 1st and 2nd principal components were plotted. The features that were close to each other are more correlated. From the biplots, we can see that 'meanPurchasePower' and ‘meanCityLevel’ have a strong correlation, ‘meanEducation’, ‘plus’, ‘MaritalStatus’, ‘meanAge’ and ‘meanUserLevel’ have a strong correlation, and ‘clickVolume’ and ‘ma14SalesVolume’ have a strong correlation. This information can help the company to improve their sales moving forward.

Other than calculating the predicted sales, a histogram is also plotted to find the distribution of the predicted demand. From the graph, exponential distribution is chosen for the predicted demand. Knowing the predicted sales and demand for the different products allows us to make better inventory decisions which will maximise the profits.