Lesson Learned Report

For Predicting Customer Preference

# Overall Impression

## Big Data and application in R

Today, most industries are facing the transition period, where the data sets are too large or complex to be dealt with by traditional data-processing software. “Big data” is a field that helps organizations harness their data and use it to identify new opportunities, which in turn, leads to smarter business decisions, more efficient operations, quicker responses, and ultimately happier customers and higher profits.

R is a programming language for statistical calculation. It is widely used among statisticians and data miners for developing statistical software, performing data analysis, and plot the graphics.

In this session of study, I learnt to use RStudio to write simple codes, optimize processes, select and train the best winning model, which eventually help us to identify customer preference, predict sales volume, and perform market basket analysis.

# Classification on the Brand Preference

## C5.0 Modelling Vs. Random Forest Modelling

Both C5.0 model and Random Forest (RF) can be applied on this classification task. I have run both modellings and selected the one with higher Accuracy and Kappa value. In this case, C5.0 modelling has a higher value.

95% of confidence level in Confusion Matrix is another key metrics for classification. Here C5.0 model is just marginal better than RF model on SurveyIncomplete dataset.

Therefore, I can safely apply the C5.0 Model on SurveyIncomplete to predict the Brand preference. The whole process is straightforward to follow.

## Ground Truth

Here comes the tricky part for this exercise. I first used postResample() command with SurveyIncomplete, the Accuracy rate was lower than 0.4. I started to question myself, is it a good model? Why the Accuracy rate is so low given that the original Accuracy rate with CompleteSurvey is over 0.92?

The underlying reason was “brand” attribute in the SurveyIncomplete contains no "truth". Brand in SurveyIncomplete contains corrupted data. There is no ground truth for comparison, therefore the low accuracy rate should be thrown away.

Accuracy and Kappa can only be found via cross validation and predict with a test set for this task. The test set, on the contrary, DOES have ground truth for brand which explains the high Accuracy rate in the 25% testing partition.

## Recommendation on Future Similar Task

Be cautious with highly skewed data

Always use the summary() function to check for the make-up and distribution for the training dataset and testing dataset. Here I noticed the brand attribute in SurveyIncomplete are heavily skewed (63 vs 4937). This is expected as the Surveyincomplete does not have the correct brand data.

Therefore, I ignored the brand values in this summary() call and replaced its value with what’s the model predicts in SurveyIncomplete file to get a more realistic value.

In short, when you see highly skewed distribution of any datasets, you need to be alerted with Ground Truth and to be careful when dealing with them.

# Multiple Regression in Predicting Sales

## Model Selection Process

In this multiple regression task, I have tried 5 algorithms: Linear Regression, Gradient Boosting Machines, Support Vector Machines (2 ways of building the model), and Random Forest. I used a 2-step selection process described below to locate the best algorithm.

First, I run all the algorithms against the various sub datasets to get a general feeling. When viewing the results, Random Forest has the best RMSE and R Square value combinations compared with other algorithms. Therefore, I removed the other algorithms from my consideration list and focused my effort on the Random Forest only.

In the second step I focused on tuning the Random Forest result with different sub dataset and mtry values. At the end, using the sub dataset newData5 with mtry = 4, I have the best result where the RMSE value is about 790.35 and R square value is 0.84. The prediction on the testSet are RMSE: 649.10 and R square: 0.88.

Although I was told to intentionally leave the 5StarReviews attribute in the dataset, I still removed this attribute due to the overfitting concern. The RMSE values are close to 0 or very low (below 300) and R square value are close to or equal to 1. This is a strong sign of overfitting. Therefore, I did not include the 5StarReviews feature in my final training.

Despite the whole selection process really took time to try each model with each subset combination, I still feel like it is quite interesting and challenging to explore the best algorithm and used it for future prediction.

## Subset the Original Dataset by Attributes

Originally, I made a mistake of removing some attributes from the readyData set by using NULL()function, the consequence was I could not restore these attributes later when I need them, I had to re-import the .csv file into the RStudio.

After watching related tutorials, I decided to subset the readyData set (removing the BestSellerRank attribute as it is NA) and partition the rest dataset by 5 different subsets, each with different feature combinations. This allows me to easily try each subset with various models and take note to compare the results.

## Recommendation on Future Similar Task

Always try caret packages first

At the beginning of the task, I forgot to use the caret package. Instead I tried writing the linear model and SVM linear model directly (I had to install the e1071 package). Later, I was told that e1071 is an older model before caret was invented. Although it is a good experiment to try and compare key testing results among various models, it is a little overwhelming to remember all the different syntaxes of various algorithms.

The beauty of caret is its simplicity as it provides an abstraction layer on top of different models to make it extremely easy to work with all the models. In many cases I can simply change the method name in the train() function to switch to a different model. The caret package contains 238 different models which covers all frequently used algorithms. Therefore, it’s always a good idea to try caret first when applying various machine learning models, as chances are, the algorithm is included in the caret package already.

# Market Basket Analysis in Predicting Sales

## Modeling Process

### Dealing with the transactional data

Unlike the previous classification and regression tasks, market basket analysis is to discover the interesting relationship (or associations) between customer’s transactions and the items they’ve purchased. These associations can then be used to drive sales-oriented initiatives such as recommender systems like Amazon frequently bought together feature.

### Interesting patterns and item relationships

I have tried 12 sets of support and confidence combination, ranging from supp= 0.1, conf= 0.8 to supp =0.001, conf=0.6. I found there were no rule if the support level (0.1-0.5) and confidence level (0.8-0.9) are both high.

I further tried the supp and conf values up and down but did not get many rules as I want until I went with supp=0.001 and conf=0.6, I have got 3969 rules and the lift value is between 2.343-17.697. I learnt that I need to try to maximize the lift value by adjusting the support and confidence value until there is no more room to increase the its value. Moreover, I ordered the lift by top 10 products and got the rhs with various monitors and keyboards, and lhs with Dell Desktop and HP laptop, etc.

### Loading the transactional dataset

I found it was not that easy as loading a regular .csv file as before. I had failed many time when importing the original transactional data into RStudio, I got a warning message like In readLines(file, encoding = encoding) :

incomplete final line found on 'ElectronidexTransactions2017.csv'

After googling the answer, I learnt it was due to the end of file character.  To get rid of the warning I did open the csv file and hit 'enter' at the end of the last line to add a blank last line and saved it, which solved the issue.

## Recommendation on Future Similar Task

Try to get as many rules as possible

Keep tuning the support and confidence value and try to get as many rules as possible for this task. You will be benefit if more rules are generated, because more products will be included in the market basket and you will notice the association rules between each transaction. Eventually, you can make recommendation to the marketing and sales team on cross-selling and whether to provide the bundle promotion if customer buy a certain product.