# Champo Market Challenge

**Page 1 to 2- Importing and cleaning the dataset**

**Exploring the Data (Page 2-4)**

**Insights:**

* The United States sells the most carpets, followed by the United Kingdom. Other countries that contribute significantly to revenue include Italy, Romania, Australia, India, Canada, and South Africa.
* Hand Tufted carpet contributed the highest to the revenue followed by Durry and the lowest being power loom jacquared.
* The “Durry” type carpets were sold the most, but it was the second highest contributor of revenue and significantly lower than “Hand Tufted” in terms of revenue, so we can say it is much cheaper in price. Because “Hand Tufted” carpet is the highest revenue generator but the number of units sold is much lower, we can say that it is on the expensive side and thus a premium quality carpet.

**Champo markets challenges**

1. **To identify the most important customers and the most important products and find a way to connect the two using suitable attributes from data and appropriate analytics models. This would help Champo Carpets recommend ideal sets of samples to customers and help increase the conversion rate.**

**Using Random Forest to find out important features**



**We can also determine the important variables by looking at the gini reduction of each variable. We can see that AreaFt has highest gini reduction and hence the most important variable.**

**Algorithms used to find ideal sets of samples to recommend for customers**

Clustering algorithm helps to better understand customers. Customer with comparable characteristics often have similar interest, thus business can benefit from this technique by creating tailored samples for each customer segment. Determine appropriate product pricing, Design an optimal distribution strategy can be the other benefits.

**Data Preprocessing for clustering:**

Data Preprocessing, We consider the “data for clustering” sheet for customer segmentation using clustering. As a first step to preprocessing, since categorical variables cannot be used in the clustering algorithms, we remove the row labels column. To be more specific, the range of categorical variables (e.g. row variables in this data) is discrete (one of the customer’s name), hence cannot be directly combined with a continuous variable and measured the distance in the same manner. Since any clustering algorithm interpret the closeness between data points based on a distance measure, it is important to reconcile all dimensions into a standard scale. An appropriate type of data transformation should be selected to align with the distribution of the data. For the case of this dataset, we standardize all the variables between 0 and 1. With the variables we have, we can divide the dataset into two subsets. One which will have the variables Sum of QtyRequired, Sum of TotalArea and Sum of Amount. The other variables will be part of another subset. This way we can cluster the similar customers in a better way. We can use principal component analysis to reduce the dimensions in the case of second subset. We’ll also use elbow method to determine the number of actual clusters.

**Clustering Algorithms (Code and output Page 50-57)**

For globular shaped clusters, center-based algorithms (K means) are more adaptable where as if the cluster are irregular in shape and have a lot of noise then density based algorithms (DBSCAN) are more applicable. Since we have reduced the dimensions of the datasets as explained in Q4 using K means clustering algorithm will give us the desired results. K means algorithm uses the Euclidean distance to interpret the closeness between data points. The number of clusters K can be found out using the elbow method.

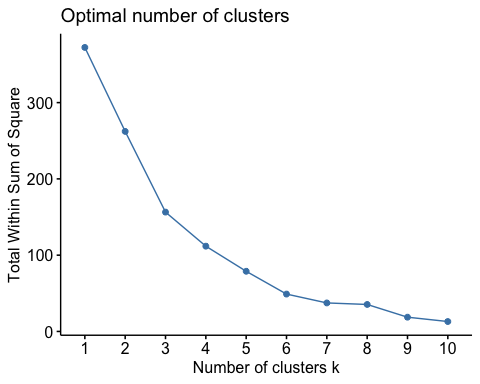
**Using K-means to solve the problem**

**Steps Followed:**

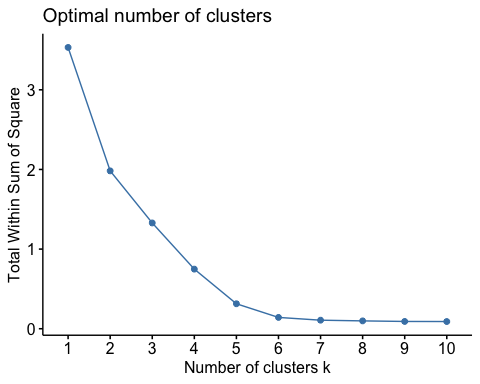
* Remove the character variable column and also standardize all the remaining columns
* Divide the dataframe to two subsets (Carpets\_sum and Carpets\_type) for k means clustering
* Use pca for carpets\_type subset to extract the principal features and discard the remaining.

**From summary it is evident that we can capture 85% of the information in the dataset (10 variables) can be encapsulated by just the first 5 Principal Components (Page 51)**

***Consider the number of clusters for the carpets type dataset***

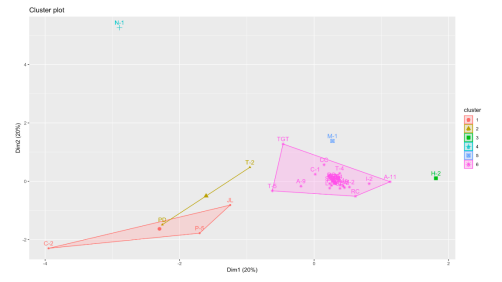


***Consider the number of clusters for the carpets sum dataset***



In both the cases we can see that there is no significant difference in the sum of squares from k value equal to 6. Hence we choose the number of clusters to be 6.

***#Applying k-means on carpets\_transform datset***



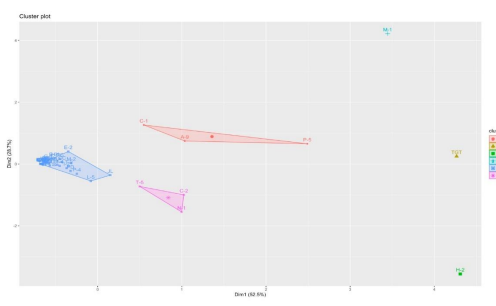
**Results:**

**The characteristics of clusters**

* Cluster 1 (C2, P5, JL) - Customers from this cluster belong to countries USA and UK. They prefer Durry, Hand Tufted and knotted carpets. We can also say that these customers prefer Chindi stripe and tikki designs.
* Cluster2 (T2, PD)- Customers from this cluster belong to the countries Belgium and Italy. They place orders in small quantity for Jacquard type carpets and they prefer rectangular type of carpets.
* Cluster 3(N1)- The customer belongs to the country USA and has ordered hand tufted carpets in a very high number thereby creating a huge revenue.
* Cluster 4 (H2)- The customer belongs to the country USA and generally prefers Durry carpets that are round in shape and have a jute design.
* Cluster 5(M1)- This customer generally prefers ordering hand woven carpets in bulk.
* Cluster 6 (rest of the customers)- Highest revenue generating customers in this belong to Australia and Brazil. Variety of items are preferred by the customers in this cluster but majorly sticking to rectangular carpets.

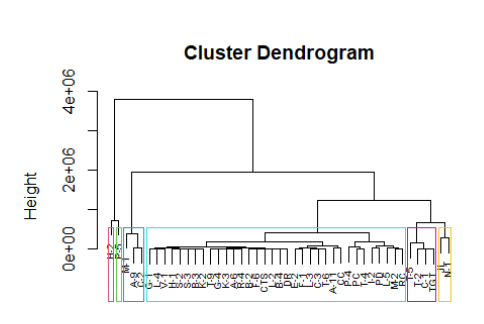
**The significant variables are ITEM NAME, Country Name and CustomerCode.**

**#Applying k-means on carpets\_transform datset**



The significant variables are sumofquantity, sumoftotalarea and sumofamount. We can see that majority of the customers belong to a single cluster in both the cases. The clustering is also very similar between the two graphs thus strengthening the confidence of it.

**Hierarchical clustering (Page 54-57)**



Using the hierarchical model, we can say that T-5, T-2, C-1 and TGT are similar to each other. Similarly, JL and N-1 are similar each other. Also, M1, A-9 and C-2 belong to the same cluster as they are similar to each other. H-2 and P-5 belong to their respective clusters. The rest of the customers belong to one individual cluster

**Recommender System (Code and output Page 57-59)**

The following recommendations can be made to the customers

Using the correlation matrix, we can see that customers N1 and C1 have a very high correlation of 97%. Therefore, after constructing the recommender matrix, we can recommend N1 to order Knotted carpets in neutral shades.

We can also see that T-5 and PD are similar as they have correlation of 99%and we can recommend Hand Tufted carpets to PD that are round and in shades of pink and blush pink.

Another set of similar customers are PC and CC as they have correlation of 99% and we can recommend handloom carpets to PC that are round and in shades of navy and blue.

1. **What are the challenges that Champo Carpets faced to reduce the wasted samples? What Machine Learning models can Champo Carpets use to solve their problems?**

Champo Carpets can use various ML algorithms to solve their problem. Champo Carpet's main aim is to reduce the number of false positives. When the order is actually not converted, but they are predicted as converted, there will be loss for Champo carpets as the samples made are wasted and it is expensive to make each sample. Hence, the champo carpets must be focusing on improving the precision (actually not converted but predicted as converted).

All the ML classification algorithms like decision trees, random Forest can be used to find out the precision of test data. When it comes to improving the precision, it is better to use random Forest than decision trees as it prevents overfitting by using multiple trees. Neural networks can also be used to predict the future data as it discovers any complex relations hidden in the data.

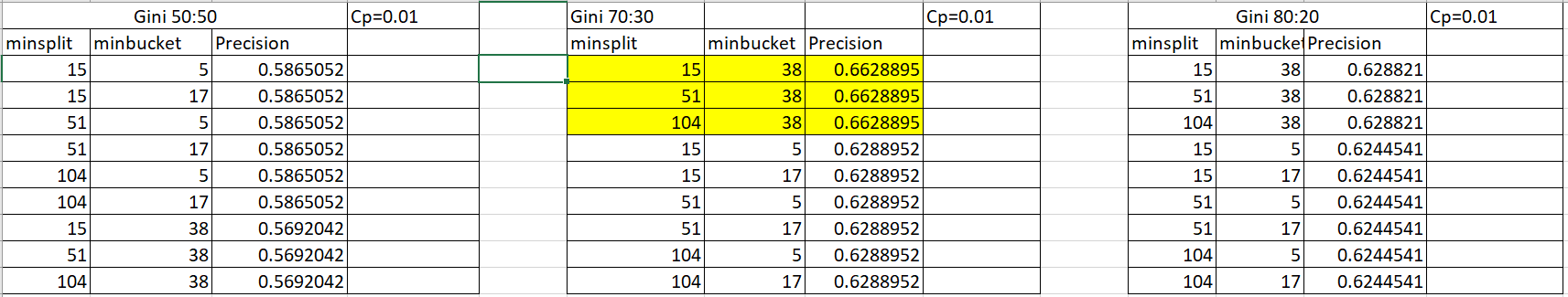
With regression, we can predict the output based on input variables. We can find out the importance of the variables by checking if they are statistically significant or not. Logistic regression can be performed to understand the relationship between predictor variables and probability of orders getting converted to samples.

**Various ML Models on Balanced and imbalanced data**

When the order is actually not converted, but they are predicted as converted, there will be loss for Champo carpets as the samples made are wasted. Hence, we chose precision as our performance metric as it is crucial to reduce the False positives(actually not converted but predicted as converted)

**On Balanced Data**

**Decision Trees (Code and output Page 5-20):**



According to the decision trees that have been constructed above, we can conclude that 70:30 split with the highlighted pruning parameters gives us the best precision.

**Random Forests (Code and output Page 20-22):**

Constructing random forests with different ntree values and finding the best mtry for each model to derive a single model with the best performance.

According to the random forest models constructed above, the random forest model1 is considered as it gives us the better precision.

Model1:

We can also determine the important variables by looking at the gini reduction of each variable. We can see that AreaFt has highest gini reduction and hence the most important variable.



Result for Model1:

Model2:

Result:



According to the random forest models constructed above, the random forest model1 is considered as it gives us the better precision.

We can also determine the important variables by looking at the gini reduction of each variable. We can see that AreaFt has highest gini reduction and hence the most important variable.

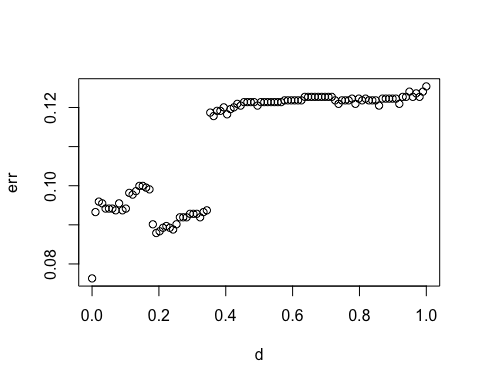
**Logistic Regression (Code and output Page 22-25)**

**Results:**

According to the P values that we have got above, we can say that the variables CountryNameBELGIUM, CountryNameCANADA, CountryNameROMANIA, ITEM\_NAMEGUN TUFTED, ITEM\_NAMEHANDWOVEN, ITEM\_NAMEKNOTTED, ITEM\_NAMEPOWER LOOM JACQUARD, TEM\_NAMETABLE TUFTED, AreaFt are significant as they are less than alpha (1%).

The anova table gives the residual deviance of null model and other variables. The more the difference in deviance between the null and residual, the best our model is doing against the null model. Hence, in the above table we can see that adding CountryName reduces the deviance. And AreaFt improves the AIC drastically and hence it is an important variable.

**Neural Network (Code and output Page 25-27):**



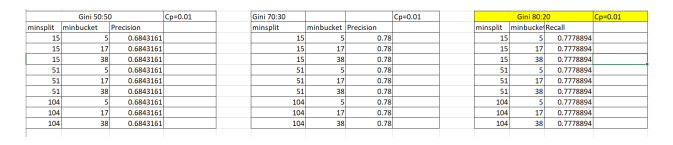
From the graph we can say that 0 is the best decay parameter as it gives the least error.

**Comparison Between Balanced and Unbalanced Data**

**Balanced Data (Balancing the data-Page 27)**

**Decision Trees (Code and output Page 27-41)**

We can observe from the precision values that we got, that balanced data gives us better precision than unbalanced data.



We can observe from the table that the 80:20 split gives us the best precision.

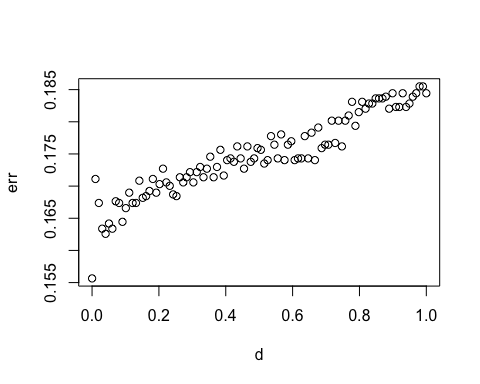
**Random Forest (Code and Output Page 41-43)**

There is a drastic improvement in precision when balanced data is used for randomforest. On balanced data, the randomForest Model2 with 300 trees gives better precision

**Logistic Regression (Code and Output Page 43-45)**

When compared to unbalanced data, balanced data gives us a greater number of significant variables. Also, the balanced data gives better AIC than unbalanced data.

**Neural Networks(Code and output Page 45-48):**

The neural network done on balanced data gives better precision of 80% than unbalanced data. 

From the graph we can see that the best decay parameter is 0 as it gives the least error.

**Final recommendation to Champo carpets would be to use all these different models as they provide different insights.**

* We can suggest them the important factors that led to conversion of orders are AreaFt, CountryName, QtyRequired as we have seen the ANOVA table of logistic regression predicting them as significant features.
* We can use the recommender systems to recommend products to customers depending on their similarity with the other customers while Kmeans can help us understand various segments of customers that we have and then make better strategies to increase their conversion rate thereby targeting the customers.
* Using Association rules, we can also suggest them the items that go well with their purchase history.
* The company can also prioritize using balanced data as it can give more accurate and precise results. - After running all the ML models above we can recommend randomForest for the Champo Carpets as it gives highest precision of 82%