

**Champo Carpets - Improving Business to Business Sales Using Machine Learning Algorithms**

Assignment 5

Big Data Technologies BIA-678

March 11, 2023



Table of Contents

1. Case Study Questions……………………………………………………………………………………………….2
   1. With the help of data visualization, provide key insights using exploratory data analysis…**………………………………………………………………………………………………………………..**3
   2. What kind of analytics and machine learning algorithms can be used by Champo Carpets to solve their problems, and in general, for value creation?……………………………………………………………………..…….………………………………………7
   3. Develop ML models to help identify features that contribute toward conversion (or non-conversion) of samples sent to customers………………………………………………………….7
   4. Discuss the data strategy for building customer segmentation using clustering. What are the benefits Champo Carpets can expect from clustering?..........................................................................................................................7
   5. Discuss clustering algorithms that can be used for segmenting Champo Carpets’ customers………………………………………………………………………………………………………………….7
   6. Develop customer segmentation using K-means clustering. Discuss the optimal number of clusters, significant variables, and cluster characteristics..............................7
      1. What are your final recommendations for Champo Carpets?........7
2. References:……………………………………………………………………………………………………………….9
3. Attachments………………………………………………………….…………………………………………………………..10
   1. ……………………………………………………………………………………………………………………………….10
   2. With the help of data visualization, provide key insights using exploratory data analysis.
      1. Some of the characteristics the study mentioned to try and understand order conversion included the following. What fibers and color trends contributed to order conversion. Color and design attributes of customers past purchases. The company also makes efforts to push samples from available raw materials. Below are examples of features and relationships that may be helpful in understanding some important relationships, in particular that may contribute to order conversion.

Map

Description automatically generated

Chart, bar chart

Description automatically generatedChart

Description automatically generated Chart

Description automatically generated Chart, bar chart, waterfall chart

Description automatically generated Chart

Description automatically generated Chart, waterfall chart

Description automatically generated

* 1. What kind of analytics and machine learning algorithms can be used by Champo Carpets to solve their problems, and in general, for value creation?
     1. Visual analytics as seen above can be helpful to begin to show what relationships exist among datapoints. In this case of Champo Carpets, we ultimately want to see what will cause a customer to convert samples sent into future orders. In this supervised situation we label our outcomes as order conversion and we can use classification algorithms to determine what features of the samples are causing a customer to convert to purchase orders. Some example machine learning algorithms can include logistic regression, decision trees, support vector machines, neural networks, etc.
  2. Develop ML models to help identify features that contribute toward conversion (or non-conversion) of samples sent to customers.
     1. Please observe python code attachment for ML models used.
  3. Discuss the data strategy for building customer segmentation using clustering. What are the benefits Champo Carpets can expect from clustering?
     1. Benefits from clustering is a way to segment customers into those who have similar purchasing or rather order conversion behaviors. So here Champo may be able to better tailor their samples to specific groups of potential customers to better improve order conversions. By creating samples that can improve order conversion to groups of customers, this will not only improve order conversion, but also save money on the number of different samples they would otherwise have made.
  4. Discuss clustering algorithms that can be used for segmenting Champo Carpets’ customers.
     1. There are several types of clustering algorithms including: K-means, hierarchical, DB scan, Spectral, etc. To discuss a few:
        1. K-means: this randomizes center points and minimizes the sum of squared distances to the center points (centroids). The center point gets reassigned based on the mean after each iteration. The number of k or centroids is predetermined. The best way to determine this is through the elbow method where there is the greatest reduction in of sum of squared distances from randomly assigned clusters.
        2. Hierarchy clusters builds a hierarchy of clusters. Data points start as a separate cluster and gradually merged into larger clusters. Alternatively, data points can start out as clusters and are recursively split into smaller clusters.
        3. DB Scan groups data points existing in high density regions of the feature space.
        4. Spectral clustering uses eigenvalues and eigenvectors of a similarity matrix to cluster data. Thus data is projected to a low-dimensional space and data in this space becomes clustered. This might be best used for image segmentation, NLP, and dimensionality reduction.
  5. Develop customer segmentation using K-means clustering. Discuss the optimal number of clusters, significant variables, and cluster characteristics.
     1. Please observe attachment for code performed for the following responses:
        1. Optimal number of clusters was found to be 3. This was based on a number of analytic components including PCA, performing the elbow method and the silhouette score.
        2. There are a number of ways the significant variables were attempted to be discovered. In order of complexity these include from simple relationships to linear to non-linear relationships:
           1. Correlation matrix demonstrates AreaFt, Other, Knotted, Belgium and USA (maybe also UK) have a relationship with order conversion
           2. Chi-square gives us AreaFt, QtyRequired and India though the third is a bit more difficult to discern
           3. Logistic Regression tells us these are AreaFt, QtyRequired and Blegium. There are a few others with high feature importance, however we’ve determined there are 3 main features. These features are determined with 80% accuracy.
           4. Decision Tree Classifier tells us with 91% accuracy that the most important features are Other, Belgium and Knotted.
           5. Random Forests however with 84% accuracy gives us AreaFt, QtyRequired and India
           6. Recursive Feature Elimination with Cross-Validation lastly tells us the most important features are USA, QtyRequired and Hand Tufted with cross-validation score of >80%.
           7. PCA however tells us that the most explained variance comes from AreaFt, QtyRequired and Hand Tufted.
        3. Chart, line chart

           Description automatically generatedChart, scatter chart

           Description automatically generatedCluster characteristics were explained by performing kmeans on the pca derived data. Without this there was very poor clustering ability with very high overlap. However after PCA we can see distinct clusters with a much higher silhouette score average 70% versus .09% the latter when performing without PCA. Specific characteristics we can see include cluster size where we see two of the three are of similar size. Additionally, we see distinct clusters with centroids in the middle and not on the boundaries. However, these clusters are very close to one another indicating there is possibly similarity between them. Two other characteristics we can look at are within-cluster and between-cluster sum of squares. However, these are seemingly arbitrary numbers and without references to other models it is difficult to distinguish if what we are seeing is good or not so good. These values can also be seen on the attachment.
     2. What are your final recommendations for Champo Carpets?
        1. Based on our analytics and machine learning algorithms we can see very distinct patterns that may help Champo make some decisions to improve order conversion meanwhile saving money on the samples they are creating. According to our algorithms it appears the most important features are the size of the sample or AreaFt and the quantity that are sent. This is consistent and clear and can be seen in our visual analytics. This perhaps makes sense because by sending a certain number of samples of different sizes a customer may be able to better see how the carpet will work with a specific room or a number of rooms. If a carpet is the wrong size it can make or break how good it looks. So by seeing the right size in a room or having enough to sample in multiple rooms, the customer may better appreciate how the carpet looks and thus create an order. The last feature is more difficult to observe both visually and through predictive means. It seems to be between country and item name. This makes sense because likely there are some countries who prefer one style versus another country. Also, some countries may appreciate a diversity of styles. So here it is necessary for Champo to detail this very closely and understand who there customers are based on their purchasing behaviors and swatches they send in to know what kind of sample style to send despite there being more popular styles for specific each country.

References:

Goyal, Priyank, et al. “Champo Carpets: Improving Business-To-Business Sales Using Machine Learning Algorithms.” *Indian Institute of Management Bangalore*, 2021, pp. 1–5. *Harvard Business Publishing Education*.