**DePaul University: College of Computing and Digital Media**

Group 5 - Capstone Project - Superstore Sales

**ABSTRACT:**

The dataset chosen for the capstone project is the Superstore Sales downloaded from the Kaggle website. The project details with various models to find meaning insights from the orders made through online. The dataset deals with the USA location of various products the main categories being furniture, office supplies and technology. Various analysis and machine learning models are implemented to study the data and gain insights by analyzing them. Along with the analysis exploring and finding patterns this paper pretty much covers overall basis in understanding and implementation on machine learning algorithms. As the data stretches from 2011 to 2014 we can say that visualization tool was used to picturize the relationship between variables. As a group project we tried to find what features were responsible for the profit as well as which relationship between variables in finding relevant insights. The project started with cleaning and pre processing of data then moving to exploration of data analysis. As many of the variables in the data are categorical transformation was performed and models were developed thus leading to evaluating the outcomes from the models.

**KEYWORDS:** Random Forest, Factor Analysis, SVM, Linear Regression, Decision Trees

# **INTRODUCTION:**

Over a decade, things have shifted gears more towards the e-commerce side. Everything has become online: groceries, electronics, stationery, accessories, and any supplies. These online services allow us to buy, deliver, return and various other services like assisting customers with video shopping to get an in-store feel. Supermarkets are frequent venues for people to purchase their necessities. This capstone project presents the goal of Superstore sales prediction for the United States.

The Superstore dataset is a publicly available dataset that simulates the sales and order data of a fictional company called "Superstore".

As part of this project, we plan to build our analysis in R as mentioned above with interactive plots to determine the profit obtained over the given years. Based on the findings we can understand which products are sold in maximum quantities with or without discounts, the state with maximum and minimum purchases and regular customers who do their shopping. These insights can help understand the demand and supply of items. Further, these can help shape the business's growth and prevent losses. The prediction of seeing the upcoming years’ profit can be estimated using this dataset.

The report will examine which products are selling well and which are not, or it will look at how sales vary by region or customer demographic. This information can be used to optimize marketing strategies or develop new products that better meet customer needs. Alternatively, the report will identify areas where costs can be reduced to increase overall profitability.

Lastly, the report will describe the process of creating and training a machine learning model on the Superstore dataset and the accuracy of the model's predictions.

# **LITERATURE REVIEW:**

**Topic 1: Predicting online shopping behaviour from clickstream data using deep learning.**

In the research paper "Machine learning for predicting online buying behaviour," a framework for forecasting potential online shopping behaviour is suggested. The framework is intended to assist online businesses in customizing their marketing approaches, maximizing their advertising campaigns, and raising customer satisfaction. In their first overview of the literature, the authors of the paper discuss how data preparation, feature selection, and model selection are all aspects of machine learning techniques for forecasting consumer behaviour. Overall, this research highlights the potential of machine learning techniques for forecasting online shopping behaviour, which can aid online retailers in streamlining their marketing plans and offering more specialized customer recommendations. Online merchants may boost sales, enhance customer pleasure, and foster enduring customer loyalty by comprehending the behaviour and preferences of their customers.

**Topic 2: A System for Classifying E-Customer Sessions Based On Support Vector Machine**

In the paper, a model for forecasting customer behavior during online purchasing is presented using the Support Vector Machine (SVM) classifier. The authors train and evaluate the model with information from online transactions as well as a number of customer-related variables like demographics, website activity, and previous purchases. The results of the algorithm demonstrate how accurately it can forecast customer behavior in online purchases. Machine learning algorithms are also capable of handling enormous amounts of data and automatically identifying patterns and trends that are challenging for humans to notice.

**Topic 3: Customer Lifetime Value and RFA for Forecasting Customer Class**

In the paper discussed, a customer class prediction model using the random forest method is offered to accurately categorize each online retail customer class when viewed from the perspective of customer lifetime value. Also, technique can help online businesses identify which client groups will need a lot of work to maintain retention strategies. Product interest will be offered when more research is conducted based on customer class and study of their preferences, improving sales and developing better interactions with potential customers. We might also urge our low-class clients to stick around to improve our retention strategy (Than Than Win &, Khin Sundee Bo).

**Topic 4: A Technique for Predicting Consumer Behavior Based on Machine Learning**

In this study, data mining technology is used to examine the traits of target groups where purchasing behavior would occur utilizing cluster analysis, decision tree analysis, and Naive Bayesian algorithm. To deliver tailored services that meet each client's specific needs and increase customer happiness, it is necessary to compare and analyze the effectiveness of various algorithms for predicting customer behavior. The decision tree model offers the biggest improvement, and it outperforms cluster analysis and the Naive Bayes model in terms of performance, according to the actual findings of target user classification and prediction (Jing LI, Shuxiao PAN, Lei HUANG &, Xin ZHU, 2019).

**Topic 5: Factor Analysis As ATool for Survey Analysis**

In this essay, elements are reduced from very large quantities to a more manageable number. Questions that are demonstrably invalid can be removed from the equation using factor analysis. Barlett's test of sphericity and sample adequacy are established in this study. The data's multicollinearity is confirmed, and a scree plot is drawn. Variable rotation is done, and the number of variables is decreased. The final findings identify crucial elements appropriate for the data as well as useful in removing unimportant variables for the investigation. The decision-makers can use the insights from this factor analysis to help them come up with a workable solution.

**Topic 6: Principal Component Analysis in Sensory Analysis**

This paper is about how principal components analysis is applied to descriptive analysis. The input is fed as a matrix format of data to the PCA and the covariance matrix or the correlation matrix is determined by the statistical methods. Based on the survey paper the PCA method was implemented that resulted with 22 correlation and 7 covariance matrix. The rest didn’t show up which method they used. In this paper covariance was chosen as the descriptors were all same. If they produce different, the data has to be considered with more attention.

**Topic 7: Predication of online sales using linear regression.**

In order to help a huge superstore, raise profits, strengthen their brand's competitiveness in accordance with market developments, and improve customer satisfaction, the goal of this paper is to evaluate and forecast their sales. For sales prediction, a well-known algorithm in machine learning called the linear regression algorithm is used. Sales data from the years 2011 to 2013 is used, and 2014 predictions are made for the numbers. Real-time data for 2014 is also gathered, and this data is contrasted with the projected data to ascertain how accurate the prediction is. This is done so that we may contrast our findings with genuine ones.

**Topic 8: Using regression trees to forecast grocery sales.**

The qualities of a point of sale's surrounding surroundings are just one of many variables that affect how well it performs. This study's main objective is to examine how the external environment of a chain of supermarkets affects the operation of the store. Data reduction of store properties is accomplished via principal components analysis (eg visibility and accessibility). For forecasting and clustering retail sales, a regression tree is employed. Smaller and larger stores are divided by the first tree branching. Sales at smaller establishments are particularly affected by factors like poor store visibility and excessive traffic. Larger points of sale perform better when there are amenities like public transportation available and no unique competitors nearby.The presented tree model has been used successfully to predict the performance of new stores, in addition to providing insight into the performance of the existing stores. This is true despite various drawbacks (such as the use of a static modeling approach).

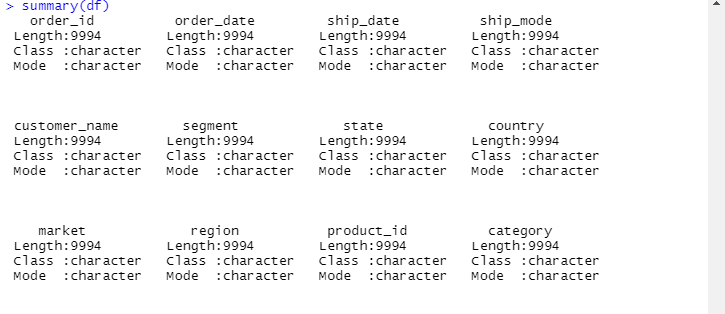
DATA:

This capstone project presents the goal of Superstore sales prediction for the United States. To give an overview of this project dataset, the data is more about details regarding online purchases. There are 9994 records with 21 columns which also include 2 date fields. The dataset contains products which are electronics, stationery supplies, furniture, accessories related to offices, storage units and many more and it is limited to the grocery.

Categorical columns- 13

Numerical columns- 6

The dataset is collected at a specific time between 2011 to 2014. So, we would like to focus on time series analysis, and visualization techniques and dig into the analysis of the data exploring variable selection. below is the summary of our dataset.



**MODELS:**

1. Random Forest Model
2. Factor Analysis
3. Support Vector Machine (SVM)
4. Linear Regression
5. Decision Tree
6. Naive Bayes
7. **RANDOM FOREST MODEL**

A popular machine learning method called Random Forest is used for classification, regression, and other applications. It uses an ensemble learning approach to mix different decision trees in order to increase accuracy and decrease overfitting. In random forest models, the relative contribution of each input variable to the prediction of the target variable is measured using a metric called variable significance. The Superstore sales dataset for the USA may be used to identify the factors that have the biggest impact on sales using variable significance, allowing businesses to direct their efforts and resources accordingly.Chart

Description automatically generated

Figure 1:

Each tree in a random forest model is constructed using a random subset of the input variables, reducing overfitting, and enhancing model accuracy. The variable relevance can be determined after the model has been trained by calculating the percentage that each input variable contributes to the model's overall prediction ability. In the case of the Superstore sales dataset, some input variables that could be included in a random forest model might include the date of the sale, the product category, the customer demographic, and the store's location. By calculating the variable importance for each input variable, the model can help businesses identify which factors impact sales most.

A prediction model for category data might be used for the Superstore dataset to determine which product category a new client is most likely to buy based on their demographic data or other pertinent attributes. By customizing their approach to certain client categories, firms can utilize this form of prediction to maximize their marketing tactics and product offerings.

The dataset must be divided into a training set and a test set to use the random forest to predict category data. The training set is used to build the random forest model, while the test set is used to evaluate the performance of the model on new, unseen data.

Chart

Description automatically generatedText

Description automatically generated

Figure 2: Random Forest Model

Chart, bar chart

Description automatically generatedTable

Description automatically generated

Figure 3: prediction using category

Next, we prepare the data by selecting the features and response variables we want to include in our model. In this case, we include the Category variable as a predictor and the Sales variable as the response.

We then train the random forest model using the random Forest function in R, setting the Category variable as a predictor and the Sales variable as the response. This model will learn how different categories influence sales and provide predictions for sales based on the category variable.

Finally, we can create a scatter plot of the predicted values using the plot function in R. This scatter plot visually displays how well the predicted sales values match the actual sales values and provides a quick overview of the model's performance.

Overall, by using a random forest model to predict sales for different segments based on the category variable, we can gain insights into how different categories influence sales and improve our ability to make accurate predictions for future sales.

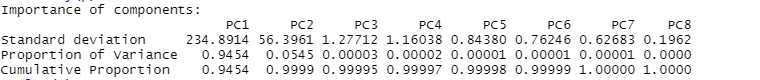
1. **FACTOR ANALYSIS**

The factor analysis is a subset of general linear algorithm. This method takes large quantity of data as input and gives us a smaller set as output. Though the data is reduced this method ensures that the data is not lost and simplified that it can easily interpreted by everyone. The main goal of factor analysis is to understand the underlying factors that are not directly observed in dataset but influence the variables. The dataset was initially cleaned and pre processed for the analysis. Exploratory data analysis were conducted. The data was then transformed to ordinal factor and numeric for model convenience. The below is the correlation plot of the numeric variables.

**Chart, bubble chart

Description automatically generated**

The correlation plot showed variables which are not correlated with any other variables. These variables were removed from the analysis. Further using the using the prcomp command the principal component analysis was performed without scaling the data.



**Chart

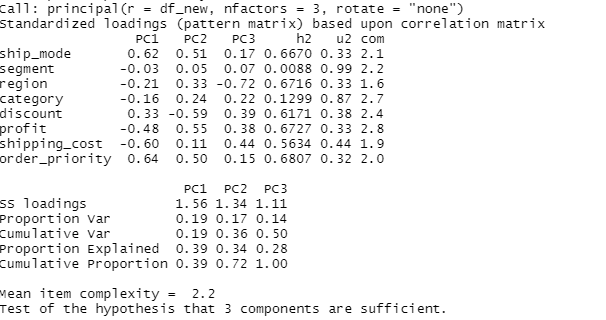
Description automatically generated**

As the data was not scaled, we can 94% of the variance is the proportion for the component. The PC1 has the highest score for profit. As PCA gives importance to variables of the component whose variance is higher. But this kind of variance distribution needs the data to be scaled.

**Chart, bar chart

Description automatically generated**

The scree plot is after the data is scaled. By doing this other PCA components will also have variance and not as before where only the first component contributed towards the variance. Shipping cost , region , profit variables contributed about 65% of the variance. PC1 and PC2 based on the region the shipping cost varies and so the delivery. Using the abline variance of 1 we see that three components contributes to the significance of the model. By performing the factor analysis we calculated the p value of <2.2e-16 which is relatively low and we can conclude that the sample is sufficient for adequacy.

****

1. **SUPPORT VECTOR MACHINE**

This is one of the most popular supervised learning algorithms which is used for classification as well as regression problems. This method applies particularly for which category has highest sales occurs and in which years. For training SVM model is by using c-classification and linear kernel trick. It is the most fundamental kind of kernel and is often of a one-dimensional kind. When there are many features, it turns out to be the best feature. For text classification issues, the linear kernel is typically favored because most of these classification issues can be divided linearly. The speed of linear kernel functions is superior to other functions. Also, which cost parameter is best for this model we can applied tune() function for better best performance value and number of support vectors.

Graphical user interface, text, application, email

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Figure: 1

From this above figure:1, we can see this SVM classification plot shown the category variables in three different colors. Also, it presents that which category has most sales occurs with their years. Here, it presents the actual values of the dataset where “Furniture” category covers the highest sell compared to the others two category.

Chart, waterfall chart

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Prediction Values** | **2011** | **2012** | **2013** | **2014** |
| Furniture | 157 | 152 | 185 | 210 |
| Office Supplies | 51 | 66 | 64 | 90 |
| Technology | 1785 | 1884 | 2331 | 3019 |

Figure: 2

For testing SVM model, we can evaluate the accuracy level by predicting the values. Above figure 2 can present the plot, and these cross tabulations present the predicted values by its number to see which year has highest sales happen also with which category product has highest sales occurs. we can clearly see that it predicts “Technology” category has 3019 products sold in year 2014, which is highest sales occurs compared to other categories.

Overall, SVM model is used to train learning data and predict the values and it gives better accuracy for our superstore sales dataset.

1. **LINEAR REGRESSION**

In a machine learning analysis, linear regression plays a key role in the evaluation of data and the establishment of a distinct relationship between two or more variables. Regression analysis measures the changes in the dependent variable when the independent variable's value changes. The data was converted to incorporate the few features that would be required to precisely anticipate overall sales and profit after all data cleaning was complete. For the dependent variable, profit, and the sales variable, we next Build a linear regression model relating Sales and Profit.

The ability to disprove the null hypothesis that there is no association between our variables will determine our ability to use our model to make predictions.

• Our data and the model match well.

**Histogram

Description automatically generated with medium confidenceText, table

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Text, letter

Description automatically generated**

Based on the result summary shown, the p-value is smaller than 0.05 as the cutoff for significance, we reject Ho . We can reject the null hypothesis in favor of believing there to be a relationship between Sales and Profit

Graphical user interface, table

Description automatically generatedChart

Description automatically generated

As has been previously mentioned, the model often underestimated the profit. This model's forecast is a cautious one.

1. **DECISION TREES**

A decision tree is a supervised machine tool for learning that may be used to classify or forecast data based on how queries from the past have been answered. The model is supervised learning in nature, which means that it is trained and tested using data sets that contain the required classification. With the help of segment, region, category, subcategory, sales, quantity, discount, and profit data as inputs, we were able to estimate the ship method.

Table

Description automatically generated with medium confidence

1. **NAIVE BAYES**

One of the most straightforward and efficient classification methods is the naive bayes algorithm, which aids in the development of quick machine learning models capable of making accurate predictions. It makes predictions based on the likelihood of an object because it is a probabilistic classifier. Even if you are working with data that has millions of data records, the preferred strategy is Naive Bayes. It is a machine learning model that is utilized for big volumes of data. When it comes to NLP tasks like sentimental analysis, it produces excellent results. Segment, region, category, subcategory, sales, quantity, discount, and profit were utilized as input in the Naive Bayes model to predict Ship Mode.

**Table

Description automatically generated with low confidence**

Given the two models mentioned above, it is clear that decision trees are more accurate than naive bayes, with an accuracy of 0.59, or 59 percent, compared to 0.56, or 56 percent. Yet, because the majority of the variables in this dataset are not connected, it cannot be classified. Further data collecting on new features is required.

CONCLUSION

We began by working on an overview of the dataset, then cleaned the dataset, worked on exploratory data analysis, and created some visualizations to show the types of variables we can use in our machine learning models. Finally, we worked with a random forest model to predict sales for various segments based on the category variable, where we can learn more about how different categories affect sales and increase our capacity for making precise predictions. For our superstore sales dataset, where we utilized the c-classification type and achieved the best performance of 0.3912323, we employed the Support Vector Machines (SVM) model, which is used to train learning data and forecast the values and it delivers greater accuracy. We can conclude that the sample is adequate for adequacy because of the factor analysis' comparatively low p value of 2.2e-16, which we obtained as a KMO. With a total MSA of.54 and three components accounting for 65% of the cumulative variation in the original dataset, linear regression is used to forecast profit from sales. The accuracy for decision trees is 0.59, or 59 percent, and the accuracy for naive bayes is 0.56, or 56 percent, which shows that decision trees have superior accuracy than the naive bayes. However, there may be certain instances with substantial errors that need to be rechecked. Yet, because the majority of the variables in this dataset are not connected, it cannot be classified. Further data collecting on new features is required. - Future research will delve further into the situations that cause significant regression model mistakes, as well as gather characteristics for classification problems.

**NEXT STEPS**

There have been implemented numerous machine learning models and analyses. We mostly worked on the superstore sales dataset in the USA. Additionally, we would like to broaden our methods globally and examine how the new data affected our predictions. If we had more time, we would like to extend our models using this dataset and machine learning techniques like parameter tuning, varimax rotation, and others..

**VISUALIZATION DASHBOARD**

Chart, bar chart

Description automatically generated

Application

Description automatically generated with medium confidence

**PROFIT BY PRODUCT**

Chart, waterfall chart

Description automatically generated

**PROFIT BY CATEGORY AND SUB-CATEGORY**

Chart, bar chart

Description automatically generated

**PROFIT BY SHIPPING**

Chart, waterfall chart

Description automatically generated

**LINE GRAPH OF DISCOUNT AND PROFIT OVER THE 3 YEARS**

Chart, line chart

Description automatically generated

**TOP 10 CUSTOMERS BASED ON THE SALES**

A picture containing calendar

Description automatically generated

REFERENCES:

1. Koehn, Dennis, Stefan Lessmann, and Markus Schaal. 2020. “Predicting Online Shopping Behaviour from Clickstream Data Using Deep Learning.” *Expert Systems with Applications* 150 (July): 113342. [https://doi.org/10.1016/j.eswa.2020.113342](mailto:https://doi.org/10.1016/j.eswa.2020.113342).
2. Suchacka, G., Skolimowska-Kulig, M., & Potempa, A. (n.d.). *CLASSIFICATION OF E-CUSTOMER SESSIONS BASED ON SUPPORT VECTOR MACHINE*. Retrieved March 21, 2023, from [https://www.scs-europe.net/dlib/2015/ecms2015acceptedpapers/0594-dis\_ECMS2015\_0120.pdf](Suchacka,%20G.,%20Skolimowska-Kulig,%20M.,%20&%20Potempa,%20A.%20(n.d.).%20CLASSIFICATION%20OF%20E-CUSTOMER%20SESSIONS%20BASED%20ON%20SUPPORT%20VECTOR%20MACHINE.%20Retrieved%20March%2021,%202023,%20from%20https:/www.scs-europe.net/dlib/2015/ecms2015acceptedpapers/0594-dis_ECMS2015_0120.pdf%20%20‌)

[‌](Suchacka,%20G.,%20Skolimowska-Kulig,%20M.,%20&%20Potempa,%20A.%20(n.d.).%20CLASSIFICATION%20OF%20E-CUSTOMER%20SESSIONS%20BASED%20ON%20SUPPORT%20VECTOR%20MACHINE.%20Retrieved%20March%2021,%202023,%20from%20https:/www.scs-europe.net/dlib/2015/ecms2015acceptedpapers/0594-dis_ECMS2015_0120.pdf%20%20‌)

1. Than Than Win1, Khin Sundee Bo, University of Information Technology, Myanmar.“Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm”. <https://ieeexplore-ieeeorg.ezproxy.depaul.edu/stamp/stamp.jsp?tp=&arnumber=9261792&tag=1.>
2. Jing LI, Shuxiao PAN, Lei HUANG, Xin ZHU. “A Machine Learning Based Method for Customer Behavior Prediction”. Technical Gazette 26, 6(2019). <https://hrcak.srce.hr/file/332434.>
3. Noora Shrestha, “Factor Analysis as a Tool for Survey Analysis”

<https://www.researchgate.net/publication/348653341_Factor_Analysis_as_a_Tool_for_Survey_Analysis>

1. Maria G. Borgognone, Javier Bussi, Guillermo Hough “Principal Component Analysis in sensory analysis”

<https://www.sciencedirect.com/science/article/pii/S0950329301000179>

1. Mastan vali, K.Sankeerthana, B.Naveen, N.Vishal “Predication of online sales using linear regression.”

[IJCRT2002169.pdf](https://www.ijcrt.org/papers/IJCRT2002169.pdf)

1. Ana Lúcia Silva & Margarida GMS Cardoso “Predicting supermarket sales: The use of regression trees”

https://link.springer.com/article/10.1057/palgrave.jt.5740150