**A screenshot of a cell phone

Description automatically generated**

**Predicting the Next Purchase**

**Natalia Yuka Hiratani**

**ID: 500999281**

**ABSTRACT**

Knowledge of consumer behaviour is crucial to the success of any business. Understanding the likelihood of a customer’s next move is invaluable. Predictive analytics can play a key role in helping supply chain management to accurately allocate product to the right place at the right time. This knowledge can contribute to outcomes such a better customer relations, an increase in sales and overall cost reduction. This project will focus on a dataset containing customer information, order details and product information. The difference between the customer’s last two purchases will be used to predict on which date, approximately, a customer will make their next purchase. The procedure will extract customer data in terms of RFM (Recency, Frequency and Monetary), duration (in days) between the last three purchases, and mean /standard deviation of the difference in time between purchases (in days). Upon completion of the feature set, the study will apply methods such as logistic regression, random forest, support vector machine, and naïve bays in order to check the accuracy of each model.

**INTRODUCTION**

Modern times have seen online shopping become the norm. In order to adapt, companies need to employ new strategies that will get their products to market, and to their customers, in the most efficient way possible. In what ways can pre-existing knowledge of customer behaviour improve product distribution? The best way is to predict what the customer’s next purchase is likely to be. In this way, companies can plan ahead and ensure customer satisfaction. The project objective is to determine on which date, a customer will make their next purchase**.** The project will use data from a supply chain company which contains customer purchase history. Machine learning techniques will be used to gauge the accuracy of the predictions.

**LITERATURE REVIEW**

Customer segmentation, which is the process of dividing customers up based on common characteristics, plays a crucial role in any business. Knowing the value of a customer is key to the development and implementation of the operational strategy (Gaurav Sharma 2018). Balancing the business’s need for efficiency with the needs of the customer is essential. Doing so provides a responsive strategy focused on speed and order fulfillment. As different customer segments have different needs, developing these strategies helps companies to reduce costs and increase revenues.

One form of customer behaviour modeling is the RFM approach. RFM stands for Recency, Frequency and Monetary. Recency refers to the last time a customer made a purchase. Frequency is how many times customers have made a purchase, and monetary is the amount a customer has spent on each purchase. RFM helps to determine a customer’s value. In other words, a customer’s worth. RFM by itself does not predict customer behaviour, but it contains information which is essential in machine learning models (Optimove 2019).

Predictive analytics is a branch of advanced analytics which is used to make predictions about unknown future events. It uses several statistical methods in order to make predictions. This project will use predictive analytics in order to predict when customers are likely to make their next purchase. This method usually uses historical data. For example, when the last purchase was made and the average time between purchases (Sam McKay CFA 2019).

By integrating machine learning methods, the customer segmentation approach, and predictive analytics, the outcome will be increased efficiency in delivery of products to market, reduced costs, increased customer satisfaction and boosts in productivity (Machine Learning in Marketing 2019).

**DATASET**

The dataset used in this project contains 53 attributes and 180.519 records. It was provided by the Mendely Data website (Source: <https://data.mendeley.com/datasets/8gx2fvg2k6/5>). Each row in the dataset corresponds to a purchase that a customer has made between 01/Jan/2015 and 31/Jan/2018. This project will use 3 attributes from this dataset: Customer\_ID, Order\_Item\_Total and Order\_Date to calculate the RFM score and create the class label necessary to apply the supervised machine learning methods.

|  |  |  |
| --- | --- | --- |
| **FIELDS** | **DESCRIPTION** | **DATA TYPE** |
| Type | Type of transaction made: CASH, DEBIT, PAYMENT, TRANSFER | Nominal |
| Days for shipping (real) | Actual shipping days of the purchased product | Discrete |
| Days for shipment (scheduled) | Days of scheduled delivery of the purchased product | Discrete |
| Benefit per order | Earnings per order placed | Continuos |
| Sales per customer | Total sales per customer made per customer | Continuos |
| Delivery Status | Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time | Nominal |
| Late\_delivery\_risk | Categorical variable that indicates if sending is late (1), it is not late (0). | Boolean |
| Category Id | Product category code | Nominal |
| Category Name | Description of the product category | Nominal |
| Customer City | City where the customer made the purchase | Nominal |
| Customer Country | Country where the customer made the purchase | Nominal |
| Customer Email | Customer's email | Nominal |
| Customer Fname | Customer name | Nominal |
| Customer Id | Customer ID | Nominal |
| Customer Lname | Customer lastname | Nominal |
| Customer Password | Masked customer key | Nominal |
| Customer Segment | Types of Customers: Consumer , Corporate , Home Office | Nominal |
| Customer State | State to which the store where the purchase is registered belongs | Nominal |
| Customer Street | Street to which the store where the purchase is registered belongs | Nominal |
| Customer Zipcode | Customer Zipcode | Nominal |
| Department Id | Department code of store | Nominal |
| Department Name | Department name of store | Nominal |
| Latitude | Latitude corresponding to location of store | Continuos |
| Longitude | Longitude corresponding to location of store | Continuos |
| Market | Market to where the order is delivered : Africa , Europe , LATAM , Pacific Asia , USCA | Nominal |
| Order City | Destination city of the order | Nominal |
| Order Country | Destination country of the order | Nominal |
| Order Customer Id | Customer order code | Nominal |
| order date (DateOrders) | Date on which the order is made | DateTime |
| Order Id | Order code | Nominal |
| Order Item Cardprod Id | Product code generated through the RFID reader | Nominal |
| Order Item Discount | Order item discount value | Continuos |
| Order Item Discount Rate | Order item discount percentage | Continuos |
| Order Item Id | Order item code | Nominal |
| Order Item Product Price | Price of products without discount | Continuos |
| Order Item Profit Ratio | Order Item Profit Ratio | Continuos |
| Order Item Quantity | Number of products per order | Continuos |
| Sales | Value in sales | Continuos |
| Order Item Total | Total amount per order | Continuos |
| Order Profit Per Order | Order Profit Per Order | Continuos |
| Order Region | Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa ,Western Europe ,Northern , Caribbean , South America ,East Africa ,Southern Europe , East of USA ,Canada ,Southern Africa , Central Asia , Europe , Central America, Eastern Europe , South of USA | Nominal |
| Order State | State of the region where the order is delivered | Nominal |
| Order Status | Order Status : COMPLETE , PENDING , CLOSED , PENDING\_PAYMENT ,CANCELED , PROCESSING ,SUSPECTED\_FRAUD ,ON\_HOLD ,PAYMENT\_REVIEW | Nominal |
| Order Zipcode | ZipCode where the order has been made | Nominal |
| Product Card Id | Product code | Nominal |
| Product Category Id | Product category code | Nominal |
| Product Description | Product Description | Nominal |
| Product Image | Link of visit and purchase of the product | Nominal |
| Product Name | Product Name | Nominal |
| Product Price | Product Price | Continuos |
| Product Status | Status of the product stock :If it is 1 not available , 0 the product is available | Boolean |
| Shipping date (DateOrders) | Exact date and time of shipment | DateTime |
| Shipping Mode | The following shipping modes are presented : Standard Class , First Class , Second Class , Same Day | Nominal |

Descriptive Statistics for numeric attributes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Days\_shipping**  **\_real** | **Days\_shipping\_**  **schedule** | **Benefit\_**  **order** | **Sales\_**  **Customer** | **Latitude** |
| **count** | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 |
| **mean** | 3.50 | 2.93 | 21.97 | 183.11 | 29.72 |
| **std** | 1.62 | 1.37 | 104.43 | 120.04 | 9.81 |
| **min** | - | - | -4,274.98 | 7.49 | -33.94 |
| **25%** | 2.00 | 2.00 | 7.00 | 104.38 | 18.27 |
| **50%** | 3.00 | 4.00 | 31.52 | 163.99 | 33.14 |
| **75%** | 5.00 | 4.00 | 64.80 | 247.40 | 39.28 |
| **max** | 6.00 | 4.00 | 911.80 | 1,939.99 | 48.78 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Longitude** | **OrderItem\_**  **Discount** | **OrderItem\_**  **Discount\_Rate** | **OrderItem\_**  **ProductPrice** | **OrderItem\_**  **ProfitRatio** |
| **count** | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 |
| **mean** | - 84.92 | 20.66 | 0.10 | 141.23 | 0.12 |
| **std** | 21.43 | 21.80 | 0.07 | 139.73 | 0.47 |
| **min** | - 158.03 | - | - | 9.99 | -2.75 |
| **25%** | - 98.45 | 5.40 | 0.04 | 50.00 | 0.08 |
| **50%** | - 76.85 | 14.00 | 0.10 | 59.99 | 0.27 |
| **75%** | - 66.37 | 29.99 | 0.16 | 199.99 | 0.36 |
| **max** | 115.26 | 500.00 | 0.25 | 1,999.99 | 0.50 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **OrderItem\_**  **Quantity** | **Sales** | **OrderItem\_**  **Total** | **Order\_Profit\_**  **Per\_Order** | **Product\_**  **Price** |
| **count** | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 |
| **mean** | 2.13 | 203.77 | 183.11 | 21.97 | 141.23 |
| **std** | 1.45 | 132.27 | 120.04 | 104.43 | 139.73 |
| **min** | 1.00 | 9.99 | 7.49 | - 4,274.98 | 9.99 |
| **25%** | 1.00 | 119.98 | 104.38 | 7.00 | 50.00 |
| **50%** | 1.00 | 199.92 | 163.99 | 31.52 | 59.99 |
| **75%** | 3.00 | 299.95 | 247.40 | 64.80 | 199.99 |
| **max** | 5.00 | 1,999.99 | 1,939.99 | 911.80 | 1,999.99 |

Descriptive Statistics for category attributes:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Type** | **Delivery\_**  **Status** | **Customer\_**  **Segment** | **Market** | **Order\_**  **Region** | **Order\_**  **Status** | **Shipping\_**  **Mode** |
| count | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 |
| unique | 4 | 4 | 3 | 5 | 23 | 9 | 4 |
| top | DEBIT | Late delivery | Consumer | LATAM | Central America | COMPLETE | Standard Class |
| freq | 69,295 | 98,977 | 93,504 | 51,594 | 28,341 | 59,491 | 107,752 |

Descriptive Statistics for object attributes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Category**  **\_ID** | **Category**  **\_Name** | **Customer**  **\_City** | **Customer**  **\_Country** | **Customer**  **\_Email** |
| count | 180,519 | 180,519 | 180,519 | 180,519 | 180,519 |
| unique | 51 | 50 | 563 | 2 | 1 |
| top | 17 | Cleats | Caguas | EE. UU. | XXXXXXXXX |
| freq | 24,551 | 24,551 | 66,770 | 111,146 | 180,519 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Customer\_**  **Fname** | **Customer**  **\_ID** | **Customer**  **\_Lname** | **Customer**  **\_Password** |
| count | 180,519 | 180,519 | 180,511 | 180,519 |
| unique | 782 | 20,652 | 1,109 | 1 |
| top | Mary | 5654 | Smith | XXXXXXXXX |
| freq | 65,150 | 47 | 64,104 | 180,519 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Customer\_**  **State** | **Customer\_**  **Street** | **Customer\_**  **Zipcode** | **Department**  **\_ID** | **Department\_**  **Name** |
| count | 180,519 | 180,519 | 180,516 | 180,519 | 180,519 |
| unique | 46 | 7,458 | 995 | 11 | 11 |
| top | PR | 9126 Wishing Expressway | 725 | 7 | Fan Shop |
| freq | 69,373 | 122 | 66,770 | 66,861 | 66,861 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Order\_City** | **Order\_Country** | **Order\_Customer\_ID** | **Order\_ID** |
| count | 180,519 | 180,519 | 180,519 | 180,519 |
| unique | 3,597 | 164 | 20,652 | 65,752 |
| top | Santo Domingo | Estados Unidos | 5654 | 52354 |
| freq | 2,211 | 24,840 | 47 | 5 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **OrderItem\_**  **Cardprod\_ID** | **OrderItem**  **\_ID** | **Order\_State** | **Order\_**  **ZipCode** | **Product\_**  **Card\_ID** |
| count | 180,519 | 180,519 | 180,519 | 24,840 | 180,519 |
| unique | 118 | 180,519 | 1,089 | 609 | 118 |
| top | 365 | 180519 | Inglaterra | 10035 | 365 |
| freq | 24,515 | 1 | 6,722 | 648 | 24,515 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Product\_**  **Category\_ID** | **Product\_**  **Description** | **Product\_Image** | **Product\_Name** |
| count | 180,519 | - | 180,519 | 180,519 |
| unique | 51 | - | 118 | 118 |
| top | 17 | NaN | <http://images.acmesports.sports>  /Perfect+Fitnes... | Perfect Fitness Perfect Rip Deck |
| freq | 24,551 | NaN | 24,515 | 24,515 |

Descriptive Statistics for boolean attributes:

|  |  |  |
| --- | --- | --- |
|  | **Late\_Delivery\_Risk** | **Product\_Status** |
| count | 180,519 | 180,519 |
| unique | 2 | 1 |
| top | TRUE | FALSE |
| freq | 98,977 | 180,519 |

Descriptive Statistics for datetime attributes:

|  |  |  |
| --- | --- | --- |
|  | **Order\_Date** | **Shipping\_Date** |
| count | 180,519 | 180,519 |
| unique | 65,752 | 63,701 |
| top | 42938.22986 | 42145.64861 |
| freq | 5 | 10 |
| first | 42,005 | 42,007 |
| last | 43,132 | 43,138 |

Using descriptive statistics, some conclusions can be made:

* Benefit\_order and Order\_Profit\_Per\_Order for each column contain the same values
* Order\_Item\_Total and Sales\_Customer x`x`for each column contain the same values
* Product\_Price and Order\_Item\_Product\_Price for each column contain the same values
* There is no data on Customer\_Email, Customer\_Password, Order\_Item\_ID, and Product\_Description.
* There are 8 null values on Customer\_Lname, 3 null values on Customer\_Zip\_Code, and 155,679 null values on Order\_ZipCode.

**APPROACH**

1. DATA EXPLORATION

This approach will use visual exploration to understand the content and characteristics of the data. This step will correct the data by reviewing its attributes, treating the outliers and filling in missing values.

Using the descriptive statistics above, some attributes of the data are able to be removed.

|  |  |
| --- | --- |
| **Attribute** | **Reason** |
| Benefit\_Order | It is duplicated. OrderProfit\_Per\_Order contains same values |
| Sales\_Customer | It is duplicated. OrderItem\_Total contains same values |
| Product\_Price | It is duplicated. OrderItem\_ProductPrice contains same values. |
| Customer\_Email | It contains the same value. |
| Customer\_Password | It contains same value. |
| OrderItem\_ID | It contains same value. |
| Product\_Status | It contains same value. |
| Product\_Descriptiton | There is no value. |
| Order\_ZipCode | There are many null values. |

1. LABEL CLASS

The class label, called “NextPurchaseDay”, is an attribute which is derived by calculating the difference between the last two purchases (in days). The Order\_Date column was used in order to check the behavioral data for each customer.

The project will also add new features to build the model. These are: the mean, and standard deviation between purchases (in days).

For customers who have made only one purchase, the “NextPurchaseDay” value was assigned as’ 60 days for a customer whose last purchase was made in 2015, 90 days for customers whose last purchase was made in 2016, and 120 days for customers whose last purchase was made in 2017.

To detect outliers in the class label, a scaling and normalization process was used. It identified 35 outliers, which represents 0.17% of the total number of customers, who were then removed.

A screenshot of a cell phone

Description automatically generated

The class label, “NextPurchaseDay”, is a continuous variable, which means that is measurable is some way. These values were then separated into 3 classes, using the quantile method. This graph shows how the classes are distributed.

A screenshot of a cell phone

Description automatically generated

Class 2 represents customers who will buy between 1 and 59 days. Class 1 represents customers who will buy between 60 and 169 and Class 0 represents customers who will buy in more than 170 days. To balance the data, an over-sampling method was applied, increasing the numbers of class 2.

1. FEATURE SELECTION

This step will calculate the RFM score for each customer and select the best attributes from the dataset.

RFM stands for Recency, Frequency, and Monetary. Recency refers to the last time a customer made a purchase. Frequency is how many times a customer made a purchase, and monetary refers to how much was spent. Customers who purchase often and spend more are considered to be more valuable customers.

For Recency, Frequency and Monetary a K-means clustering was applied for each one. K-means is an unsupervised machine learning technique that divides each customer in the *k* cluster. For each cluster, the RFM score can be calculated by adding the combined recency, frequency and monetary clusters. The Elbow Method, which is a method to help in finding the appropriate number of clusters in a dataset, was used to define the best number of clusters. The picture below shows that 3 clusters are optimal.

A screenshot of a cell phone

Description automatically generated

Customers who have 6 as an overall score, are the best customers, because on average, they bought more recently, frequently, and spend more money.



To select the best attributes, the dataset was divided into 2 parts: numerical and categorical attributes. The backward elimination process, where all the independent variables are entered into the equation first and each one is deleted, using Logistic Regression was applied, because the training and test scores for Logistic Regression are most similar. After the backward elimination process was applied, 15 attributes were selected.

|  |  |  |  |
| --- | --- | --- | --- |
| **Numerical** | **Random Forest** | **Linear Regression** | **Logistic Regression** |
| **Training Score** | 1 | 0.1893 | 0.5174 |
| **Test score** | 0.5299 | 0.1603 | 0.49903 |
| **Diff** | **0.4701** | **0.029** | **0.01837** |
|  |  |  |  |
| **Categorical** | **Random Forest** | **Linear Regression** | **Logistic Regression** |
| **Training Score** | 0.79318 | 0.0223 | 0.43502 |
| **Test score** | 0.33826 | -0.02963 | 0.41589 |
| **Diff** | **0.45492** | **0.05193** | **0.01913** |

1. Model Data

In this step, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine were used in order to model the data. For each model, a kfold cross-validation was applied, with k=10. To compare the models, Accuracy and ROC (Receiver Operator Characteristic) were used to select possibilities for optimal models.

The accuracy of a measurement is determined by how close a result comes to the true value. ROC is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold varies.

In Accuracy, the average of each interaction is shown below:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Model** | **Mean** | | Logistic Regression | 0.526193 | | Naïve Bayes | 0.450099 | | Random Forest | 0.620406 | | Support Vector Machine | 0.539202 | | A screenshot of a cell phone  Description automatically generated |

Comparison the ROC with Class 0 VS Class 1 and 2.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Comparison the ROC with Class 1 VS Class 0 and 2.

A close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Comparison the ROC with Class 2 VS Class 0 and 1.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**RESULTS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **RFM SCORE** | **# CUSTOMERS** | | 0 | 780 | | 1 | 934 | | 2 | 9,430 | | 3 | 3,691 | | 4 | 2,509 | | 5 | 1,848 | | 6 | 1,449 | | According to the table, 28% of customers have a good score, which means that they are buying frequently, more often, and spending more money. However, 54% have the opposite behaviour. |
| A screenshot of a cell phone  Description automatically generated | The histogram to the left shows that most customers have a long interval between purchases. The average is 194 days. |
| |  |  | | --- | --- | | **Model** | **Mean** | | Logistic Regression | 0.526193 | | Naïve Bayes | 0.450099 | | Random Forest | 0.620406 | | Support Vector Machine | 0.539202 |   A screenshot of a cell phone  Description automatically generated | Random forest method provides the most accurate measure between the models: 62%. Furthermore, the ROC method shows that the Class 0, in comparison with the other classes (1 and 2), is 74% accurate, which means that customers have a greater likelihood to buy infrequently. In order to improve the accuracy and the ROC in class 2, other strategies need to be implemented. |

**CONCLUSIONS**

This project shows that knowing the interval between purchases through behavioral data is a crucial point of reference for businesses. Applying machine learning techniques can be valuable tools in assisting human decision making.

**REFERENCES**

Optimove 2019, *Customer Behaviour Modelling*, Optimove, viewed November 2019, <<https://www.optimove.com/resources/learning-center/customer-behavior-modeling>>.

Gaurav Sharma 2018, *4 Types of Customer Segmentation All Marketers should know*, Gaurav Sharma, viewed 22 September 2018, <<https://www.business2community.com/customer-experience/4-types-of-customer-segmentation-all-marketers-should-know-02120397>>.

Alex Bi 2019, *Predict When Your Customers will Reorder*, Alex Bi, viewed 14 January 2019, <<https://www.klaviyo.com/blog/expected-date-of-next-order>>.

Sam McKay CFA 2019, *Predicting When Will Your Customers Purchase Next w/Power BI*, viewed 27 September 2019, <<https://blog.enterprisedna.co/predict-when-customers-will-purchase-next-advanced-power-bi-dax/>>.

Machine Learning in Marketing 2019, *Machine Learning in Marketing*, Optimove, viewed December 2019, <<https://www.optimove.com/resources/learning-center/machine-learning>>.

What is predictive analytics n.d., *What is predictive analytics?,* <<https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/>>.