Customer Segmentation and Sales Forecasting

KISHOR KUMAR SRIDHAR

17 MAY 2021

Outline of the presentation

- Data
- Data Pre-processing
- Feature Engineering
- Exploratory Data Analysis (EDA)
- Customer Segmentation using RFM Analysis
- Marketing strategies for customer segments
- Customer Segmentation using K-Means Clustering
- Sales Forecasting using Time Series Modeling (Fb Prophet Model)
- Sales Forecasting using Time Series Modeling (SARIMAX Model)
- Model comparison
- What this means for Hertz
- Conclusion

Data

The data set contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

There are 541909 rows and 8 columns in our data

Data | Columns

1. **CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.**
2. **InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.**
3. **StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.**
4. **Description: Product (item) name. Nominal.**
5. **Quantity: The quantities of each product (item) per transaction. Numeric.**
6. **InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.**
7. **UnitPrice: Unit price. Numeric, Product price per unit in sterling.**
8. **Country: Country name. Nominal, the name of the country where each customer resides.**

Data Pre-Processing

- **Removing “Quantity” column values that are less than 0.**
  (Assuming that the “Quantity” ordered can’t be less than 0)
- **Removing “UnitPrice” column values that are less than 0 .**
  (Assuming that the “UnitPrice” can’t be less than 0)
- **Removing the NULL values in the “CustomerID” and “Description” columns**
- **Removing duplicate rows**

Before Data-Preprocessing: **541,909** rows and 8 columns.

After data-preprocessing: **392,732** rows and 8 columns.

Feature Engineering

- **Creating a new column named “Sales” by multiplying “Quantity” and “UnitPrice” columns**
- **Splitting the “InvoiceDate” into the following features**
  - Date of the order
  - Year of the order
  - Month of the order
  - Week of the year the order was placed
  - Day of the week the order was placed
  - Hour of the day the order was placed
- **Creating sessions - 'Morning’, 'Afternoon’, 'Evening’, 'Night’**

After the data manipulation and feature engineering, our data now has **392,732** rows and **17** columns

Exploratory Data Analysis

- Count of transactions for each country
- Count of transactions, quantity of orders, and sales on each day of the year
- Quantity of orders and sales on each month of the year
- Sales on each week of the year
- Sales on each day of the week
- Count of orders on each hour of the day
- Count of orders on each session of the day (Morning, Afternoon, Evening, Night)
- Top 20 orders based on the description

Customer Segmentation

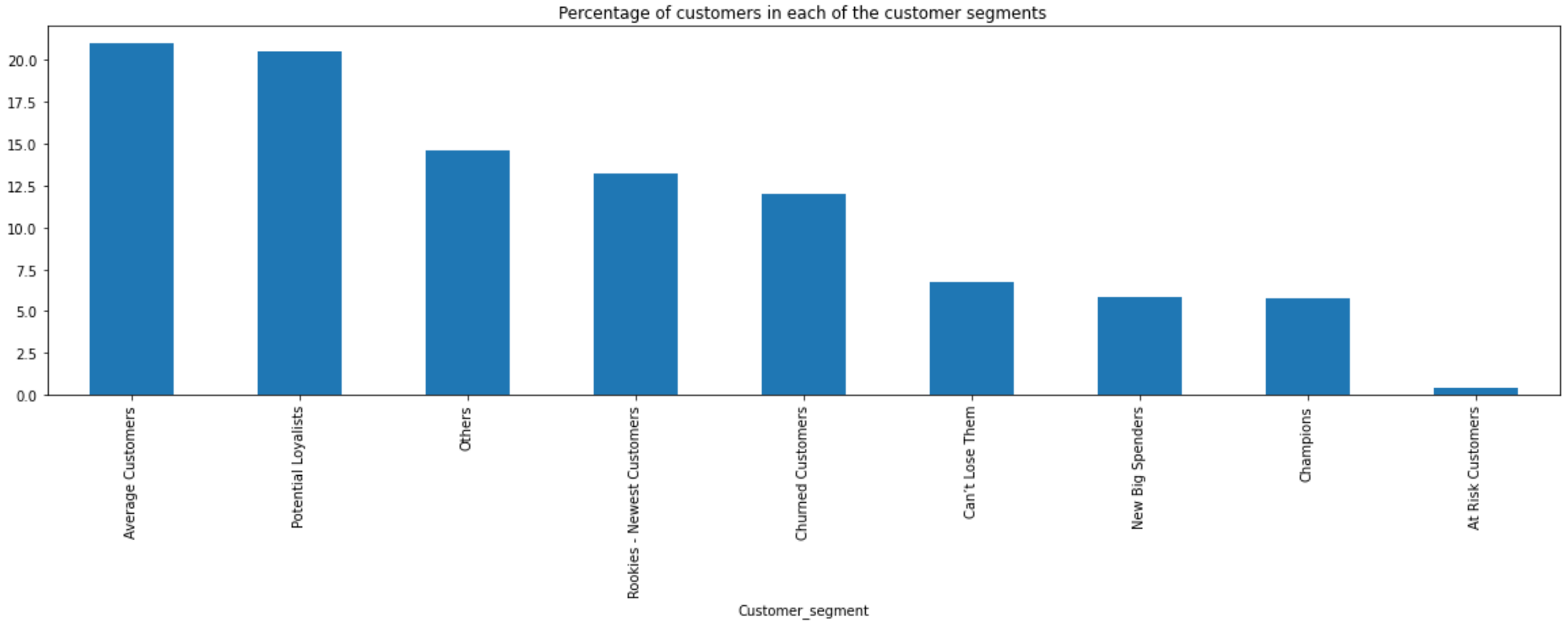
RFM Analysis

The RFM stands for

- **RECENCY (R):** Days since last purchase
- **FREQUENCY (F):** Total number of purchases
- **MONETARY VALUE (M):** Total transaction value that a customer spent

The idea is to segment customers based on when their last purchase was, how often they've purchased in the past, and how much they've spent overall.

Customer Segments



Customer Segments

- Champions (5.78% of the customers)
- Potential Loyalists (20.48%of the customers)
- Average Customers (20.99% of the customers)
- Rookies - Newest Customers (13.20% of the customers)
- Churned Customers (12% of the customers)
- Can't Lose Them (6.75% of the customers)
- New Big Spenders (5.83% of the customers)
- At Risk Customers (0.36% of the customers)
- Others (14.5% of the customers)

Champions

Champions (5.78% of the customers):

RFM Score: 333

Who They Are:

- Highly engaged customers who have bought the most recent, the most often, and generated the most revenue.
- Potential high-valued customers.
- Generate a disproportionately high percentage of overall revenues
- Focusing on keeping them happy should be a top priority.

Marketing Strategies:

- Communications with this group should make them feel valued and appreciated. We need to reward these customers.
- They can become early adopters for new products or new feature releases of the app and will help promote the brand. Suggest them to share your products with their friends or family using "Referral Program". It will help to increase conversion rates.
- Focus on loyalty programs for them. These customers have proven to have a higher willingness to pay, so instead of using discount pricing to generate incremental sales, we should focus on value added offers through product recommendations based on previous purchases.
- Further analyzing their individual preferences and affinities will provide additional opportunities for even more personalized messaging.

Potential Loyalists

Potential Loyalists (20.48%of the customers):

RFM Score: 322, 323, 332, 223, 233

Who They Are:

- The customers in these RFM groups are our recent customers with average frequency and who spent a good amount.

Marketing Strategies:

- This group has the potential to become Champions.
- Highly promising customers and need to taken care of by offering annual or quarterly membership programs to them with additional benefits
- Loyalty programs or recommend related products to upsell them may help them become our Champions. Loyalty programs are effective for these repeat visitors.
- Advocacy programs and reviews are also common strategies.
- Lastly, consider rewarding these customers with Free Shipping or other like benefits. By doing so, they will shop more frequently and for more amount.

New Big Spenders

New Big Spenders (5.83% of the customers):

RFM Score: 313, 312

Who They Are:

- These customers are the ones who have a high overall RFM score based on most recent visits and high spending but are not frequent shoppers.

Marketing Strategies:

- Start building relationships with these customers by providing onboarding support and special offers to increase the frequency of their visits.
- Potentially high-valued customers.
- Since they have purchased very recently with high spending amount, as long as we increase their frequency, they will become the best customers with high loyalty. They need to be cultivated by the brand to become long-term loyalists.
- We can welcome them with personalized email with a coupon code to encourage repeat purchases. We can also enroll these customers into a loyalty or membership program that rewards order frequency.

Average Customers

Average Customers (20.99% of the customers):

RFM Score: 212

Who They Are:

- These customers have purchased fairly recently, and their spending is average, but the purchase frequency is relatively low.

Marketing Strategies:

- Considering the huge customer base, this group can have large potential values and we need to incentivize them to spend more.
- For example, we can reward them with special discount offers, free shipping or other benefits.

Rookies - Newest Customers

Rookies - Newest Customers (13.20% of the customers):

RFM Score: 211, 311

Who They Are:

- Customers who visited the website recently often, but do not frequent it or spend a lot.
- Customers who look at articles in the products without making any purchases.

Marketing Strategies:

- These customers visit relatively often but only to look at the products and not buy them. Even if they do end up buying, they are only interested in the low-cost items.
- Hence, providing them discounts at a lower cost to encourage their buying is the way to attract these customers.
- Having clear strategies in place for first time buyers such as triggered welcome emails will pay dividends.

Can’t Lose Them

Can’t Lose Them (6.75% of the customers):

RFM Score: 222

Who They Are:

- These are customers who used to visit and purchase quite often and spend a decent amount of money.

Marketing Strategies:

- The main aim is to keep them coming back to the website for more purchases.
- We may run surveys to find out what could be improved about their experience
- Avoid losing them to a competitor by providing timely offers on relevant products and free-shipping benefits.

At Risk Customers

At Risk Customers (0.36% of the customers):

RFM Score: 133, 123

Who They Are:

- Great past customers who haven't bought in awhile. They are your customers who purchased often and spent big amounts but haven’t purchased recently.

Marketing Strategies:

- Customers who are in the risk of churning and we must bring them back with relevant promotions.
- Customers leave for a variety of reasons but one of the major advantages of these customers is that they have frequented the website many times in the past, so we have a good record of the product they are interested in.
- Send them personalized reactivation campaigns to reconnect and offer renewals discounts and suggest helpful products to encourage more recent visits to the website.
- Suggest "Referral Program" and "Annual Membership Program" to prevent these customers from churning as they were once frequent and high spenders in the past.

Churned Customers

Churned Customers (12% of the customers):

RFM Score: 111

Who They Are:

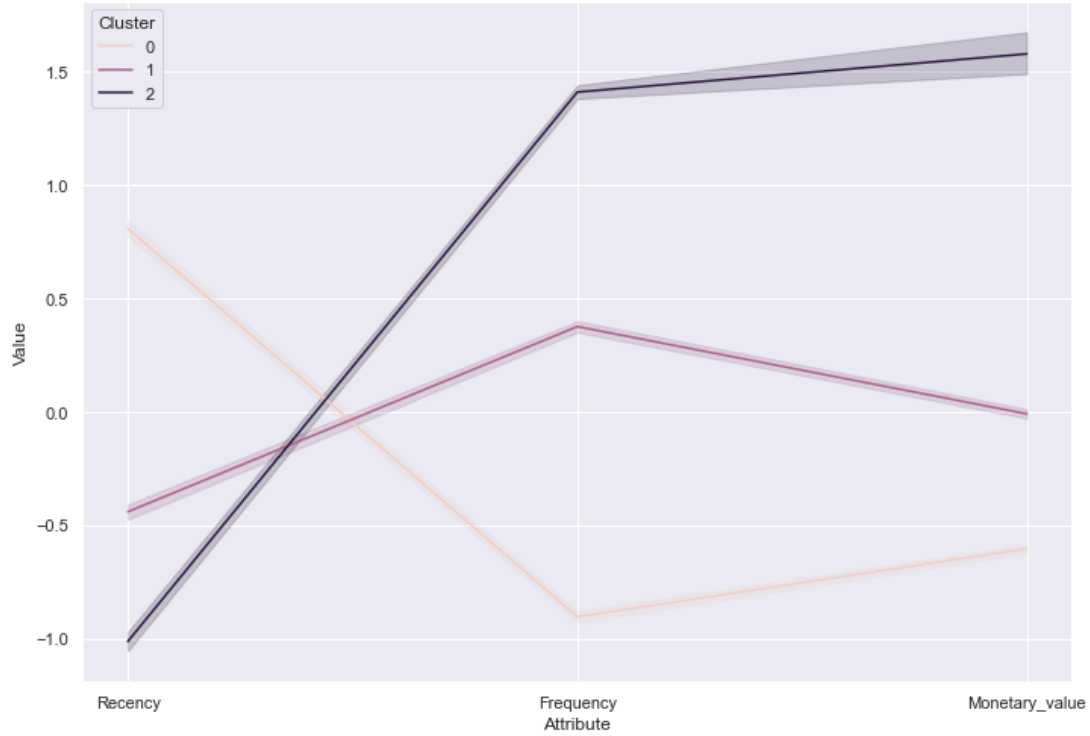
- They probably bought once or very few times and they bought for very less amount.

Marketing Strategies:

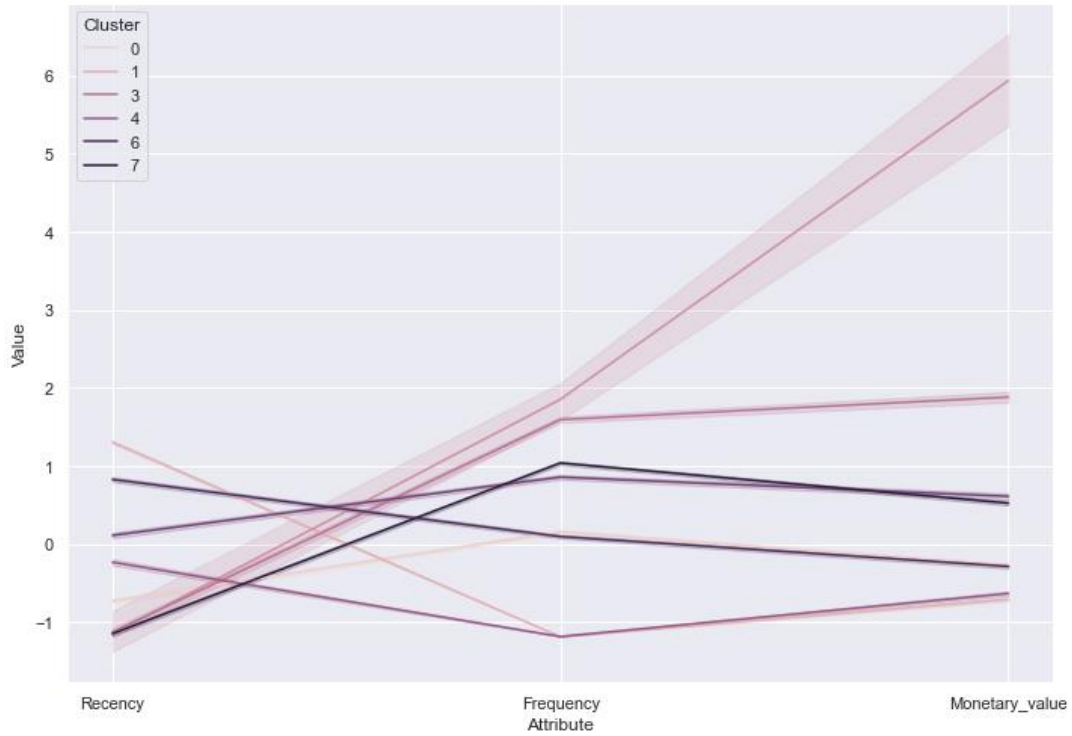
- Give lost customers an incentive to come back. We can send this campaign directly to inactive customers with tagline phrases like "Haven’t seen you in a while", "Let’s catch up", "Come back and receive [promotion details]" which would encourage them to come back.
- Another way is to address common complaints through a social media campaign and create visibility of a productive change in process.

Customer Segmentation using K-Means Clustering

K-Means clustering with 3 Clusters



K-Means clustering with 8 Clusters



With the help of specific domain knowledge, we may be able to interpret these clusters better for identifying customer segments.

Sales forecasting

Sales Forecasting using Time Series Modeling (FB Prophet Model)

FB Prophet model:

- FB Prophet is a procedure for forecasting time series data based on a Generalized Additive Model (GAM) where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
- The systematic component of an additive model is the arithmetic sum of the individual effects of the predictors. Prophet combines seasonality, trend, and holidays.

$$y(t) = g(t) + s(t) + h(t) + et$$

- g(t) = trend component (the trend function which models non-periodic changes in the value of the time series)
- s(t) = seasonal component (represents periodic changes (e.g., weekly and yearly seasonality))
- h(t) = holiday component (represents the effects of holidays)
- et = remainder component (represents any idiosyncratic changes or errors)

Evaluation metrics

- **Mean Absolute Error (MAE)** : The Mean Absolute Error is the average of the absolute difference between the actual and predicted values.
- **Mean Absolute Percentage Error (MAPE)** : The Mean Absolute Percentage Error is the average of absolute percentage errors
- **Mean Squared Error (MSE)** : Mean Squared Error represents the average of the squared difference between the original and predicted values
- **Root Mean Squared Error (RMSE)** : Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals

We shall use MAPE to evaluate our model. Because MAPE is robust to outliers.

Model Performance

	Base Model	Improved Model
MAE	19,395.0	10,953.0
MAPE	44.0 %	26.0 %
MSE	526,025,768.0	207,326,231.0
RMSE	22,935.0	14,399.0

We shall use MAPE to evaluate our model. Because MAPE is robust to outliers.

Interpretation of the model results

- The Mean Absolute Percentage Error (MAPE) is reduced to 26% from being 44% in the base model. This indicates that **overall, the predicted values for the sales in the retail store, we are out with an average of 26% from the true value.**
- The Root Mean Squared Error (RMSE) is reduced to 13910.0 from being 22935.0 in the base model. This means that in general, **the model's predicted sales for the retail store is generally about 13910.0 Sterling's off.**

With the help of specific domain knowledge, we may be able to improve these scores.

Sales Forecasting using Time Series Modeling (SARIMAX Model)

Model Performance

	SARIMAX Model
MAE	15243.0
MAPE	35.0 %
MSE	637476831.0
RMSE	25248.0

With the help of specific domain knowledge, we may be able to improve these scores.

Model comparison

	FB Prophet Model	SARIMAX Model
MAE	10,953.0	15243.0
MAPE	**26.0 %**	**35.0 %**
MSE	207,326,231.0	637476831.0
RMSE	14,399.0	25248.0

We can see that our FB Prophet model performed better with the MAPE score of 26% as compared to the SARIMAX model with the MAPE score of 35%.

What this means for Hertz

Business Recommendation: Fleet Utilization

Rental Car Reservation Demand Forecast: Develop a forecast model to predict demand of car rentals by hour at each location using FB Prophet time series algorithm.

We could use external data sources such as,

- Demographics (as Regressor component)
- Weather (as Regressor component)
- Airport schedules (as Regressor component)
- Hotel occupancy (as Regressor component)
- Specific holidays (as holiday component)
- Long weekends (as changepoints)
- Social events (as changepoints)
- Post-COVID Promotional offer dates(as changepoints)

Conclusion

- We analyzed online retail dataset to perform **customer segmentation using RFM analysis**. We segmented the customers into **8 major categories** as Champions, Potential Loyalists, Rookies - Newest Customers, At Risk Customers, Average Customers, New Big Spenders, Can't Lose Them, Churned Customers and discussed potential marketing strategies for each of the customer segments. We implemented **K-Means clustering** ML algorithm that could be used with RFM scores to perform customer segmentation.
- We then used the sales data to create an **FB Prophet model** and **SARIMAX model** to forecast future sales and explore the hyperparameters we can tune to get the most out of the historic sales data. We found that our **FB Prophet model works better than the SARIMAX model.**

In conclusion, by coupling our analysis with the insights from some domain knowledge experts, we would be able to make the best out of our data.

References: I referred articles on customer segmentation from websites such as Medium, Analytics Vidhya, Stack Overflow. Used FB Prophet documentation and SARIMAX documentation for time series forecasting.

THANK YOU!!!

- FOR YOUR VALUABLE TIME -