CROP CARE PROJECT

* Business Problem:

1. To predict credit worthiness of Dealers
2. To Forecast Sales of Products

* Datasets

1. Sales of Products
2. Returns of Products
3. 2016-2018 sales Values

* EDA:

1. Sales

Dimension: 9335 \* 9

Description numeric columns:

|  |  |  |  |
| --- | --- | --- | --- |
| Columns | Mean | Std | Median |
| Quantity | 96.191 | 251.179 | 40.00 |
| Value | 4.471e+04 | 1.333e+05 | 1.712e+04 |

Description categorical columns(excluded voucher no & ref):

|  |  |  |  |
| --- | --- | --- | --- |
| Columns | Unique | Top | Frequency |
| Dealer | 675 | Del160 | 260 |
| Area | 515 | 0 | 347 |
| Products | 178 | XPLODE 250ml | 735 |

Date column

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Unique | From | To |
| Date | 291 | 2017-04-01 | 2018-03-30 |

* Outliers in ‘Value’ column

1. There is 13 missing values,
2. It have 48 entries with Value = 1
3. It have 28 entries with Value = number of Quantity , which does not make sense.

* Imputing with Median values

Checking whether unit price of single product is equal across all area wise, dealer wise and date wise

For product KHANJARBIOSL- 250 ML the unit price remains same

Index price Dealer\_id Area

18 107.0 Del1337 0

123 107.0 Del1222 KUPPAM

307 107.0 Del1142 JAGGAMPETA

383 107.0 Del164 BHIMADOLE

448 107.0 Del1616 RAYACHOTY

483 107.0 Del118 RAYACHOTY

484 107.0 Del118 RAYACHOTY

567 107.0 Del1515 NARASARAOPET

569 107.0 Del1515 NARASARAOPET

712 107.0 Del1229 V KOTA

731 107.0 Del112 V KOTA

743 107.0 Del112 V KOTA

757 107.0 Del112 V KOTA

For Product ‘XPLODE - 250 ML (1X40)’

Index unit\_value Dealer\_ID Area Date

4006 675.00 Del19 NELLORE 2017-04-03

2152 675.00 Del1317 G.V. SATRAM 2017-04-06

5188 675.00 Del1621 KAZIPET 2017-04-06

4332 675.00 Del165 PICHATOOR 2017-04-06

8550 675.00 Del1168 PUNGANUR 2017-04-07

3585 675.00 Del129 PEARU 2017-04-07

3230 357.18 Del161 TSINV CTRL 2017-04-18

3232 357.18 Del161 TSINV CTRL 2017-04-18

3229 357.18 Del161 TSINV CTRL 2017-04-18

3231 357.18 Del161 TSINV CTRL 2017-04-18

3780 534.60 Del1392 VEERAVARAM 2018-03-03

7443 534.60 Del1246 GUDUR 2018-03-05

3073 213.84 Del160 TS 2018-03-05

8279 534.60 Del121 KUPPAM 2018-03-21

3141 299.25 Del160 TS 2018-03-22

Conclusoin :

From above we can impute or correct the missing, irrelevant values in ‘Value’ columns by multiplying this unit price with number of quantities for each products. But Story is not same in other products eg ‘XPLODE - 250 ML (1X40)’ unit price is giving information about how most frequent Dealers are getting discount.

We can use this fluctuation in further analysis. So final decision is removing outlier values.

There are irrelevant features in dataset which is also removed voucher, voucher\_ref and voucher\_no.

1. Returns

* Description of numeric values;

|  |  |  |  |
| --- | --- | --- | --- |
| Columns | Mean | Std | Median |
| Quantity | 43.142322 | 52.496087 | 37.000000 |
| Value | 18044.241695 | 22937.659746 | 12847.500000 |

* Description categorical columns(excluded voucher no & ref):

|  |  |  |  |
| --- | --- | --- | --- |
| Columns | Unique | Top | Frequency |
| Dealer | 162 | Del121 | 35 |
| Products | 114 | MADALI-100GRX40 | 59 |

* Date column

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Unique | From | To |
| Date | 68 | 2017-04-22 | 2018-03-30 |

Removing irrelevant features new\_voucher, new\_voucher\_ref and new\_voucher\_no.

Top five products returned are:-

Name count

1. MADALI-100GRX40 59
2. SHAKTI- 1 LTR (1X10) 55
3. EFFECTP- 250 ML (1X40) 51
4. KHAJANRBIO-SP 150 GM (1X40) 48
5. XPLODE - 250 ML (1X40) 40

* There are some products which are not present in Sales dataset for example the most returned product above MADALI-100GRX40 and EFFECTP- 250 ML (1X40).
* Where XPLODE is most sales product in sales dataset.

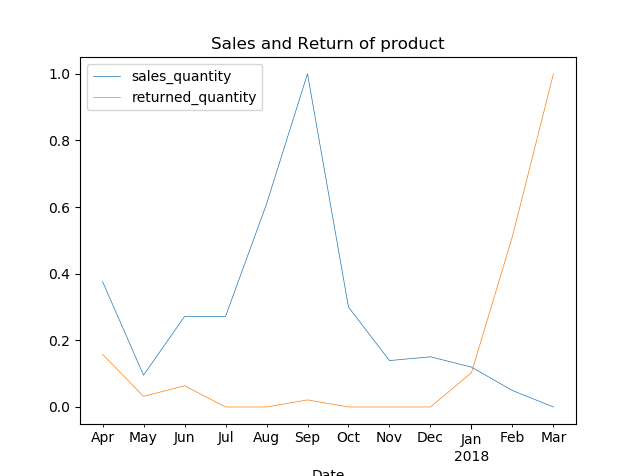
Sales VS Returns:

Sales and Returns Visualization of Two most frequent products, by taking sum of ‘Quantity’ monthly.

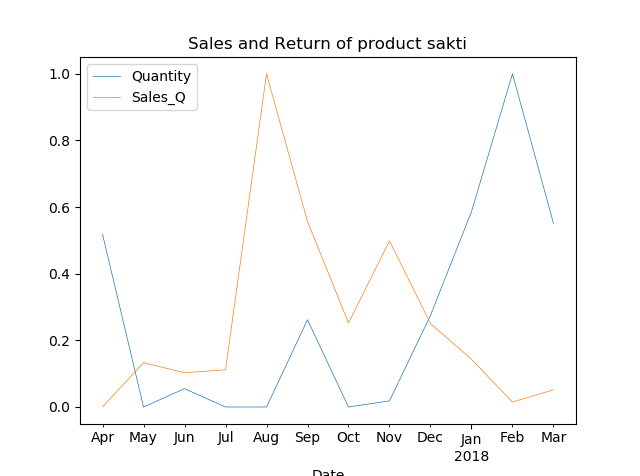
1. KHANJARBIO -SP 150 GM (1X40)

Data

|  |  |  |
| --- | --- | --- |
| Date | S\_quantity | R\_quantity |
| 2017-04-30 | 0.376556 | 0.158174 |
| 2017-05-31 | 0.095563 | 0.031847 |
| 2017-06-30 | 0.271500 | 0.063694 |
| 2017-07-31 | 0.271500 | 0.000000 |
| 2017-08-31 | 0.609029 | 0.000000 |
| 2017-09-30 | 1.000000 | 0.021231 |
| 2017-10-31 | 0.299627 | 0.000000 |
| 2017-11-30 | 0.139301 | 0.000000 |
| 2017-12-31 | 0.150552 | 0.000000 |
| 2018-01-31 | 0.119612 | 0.102972 |
| 2018-02-28 | 0.049293 | 0.512208 |
| 2018-03-31 | 0.000000 | 1.000000 |



1. For SHAKTI- 1 LTR (1X10



From above analysis we visualize that returns of products is depended on sales. As there is drop in sales return of product increases.

RFM Analysis

Instead of analysing the entire customer base as a whole, it’s better to segment them into homogeneous groups, understand the traits of each group, and engage them with relevant campaigns rather than segmenting on just customer age or geography.

**RFM factors illustrate these facts:**

* the more recent the purchase, the more responsive the customer is to promotions
* the more frequently the customer buys, the more engaged and satisfied they are
* monetary value differentiates heavy spenders from low-value purchasers

Below is the python code to calculate the Recency, Frequency, and Monetary.

rfm = Sales.groupby('Dealer\_ID').agg({'Date': lambda d: (Sales.Date.max() - d.max()).days,

'Value': lambda price: price.sum()})

rfm['frequent\_days'] = Sales.groupby('Dealer\_ID').agg({'Date': lambda d: d.nunique()})

R,F and M is rated between 1-5, where 5 is best and 1 for lowest hence dealers with 555 as RFM\_score is "the best Dealer".

I used pd.qcut() for quantile segments from 1-5

rfm['q\_recent'] = pd.qcut(rfm.recent\_days,5,[5,4,3,2,1],duplicates='drop')

rfm['q\_frequency'] = pd.qcut(rfm.frequent\_days,6,[1,2,3,4,5],duplicates='drop')

rfm['q\_monetary'] = pd.qcut(rfm.total\_spend,5,[1,2,3,4,5])

Below are names of 10 segments based on dealers Recency and frequency.

* Champions: Bought recently, buy often and spend the most
* Loyal Customers : Buy on a regular basis. Responsive to promotions.
* Potential Loyalist: Recent customers with average frequency.
* Recent Customers: Bought most recently, but not often.
* Promising: Recent shoppers, but haven’t spent much.
* Customers Needing Attention: Above average recency, frequency and monetary values. May not have bought very recently though.
* About To Sleep: Below average recency and frequency. Will lose them if not reactivated.
* At Risk: Purchased often but a long time ago. Need to bring them back!
* Can’t Lose: Them Used to purchase frequently but haven’t returned for a long time.
* Hibernating: Last purchase was long back and low number of orders. May be lost.

R M F RMF Segments

2 1 1 211 hibernating

5 1 1 511 new customers

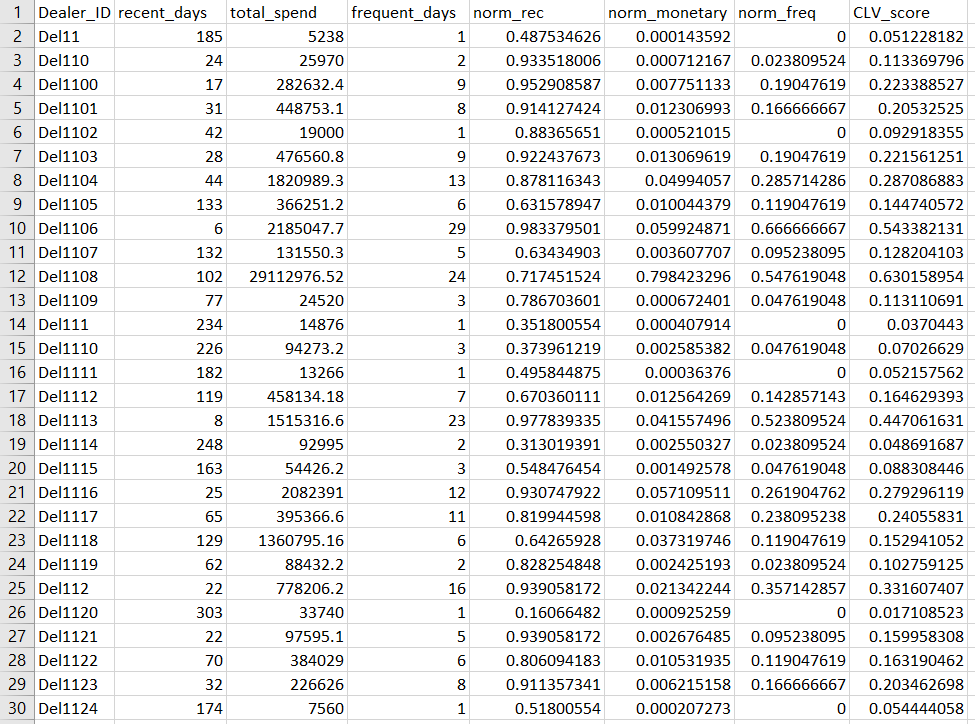
5 4 4 544 champions

5 4 4 544 champions

5 1 1 511 new customers

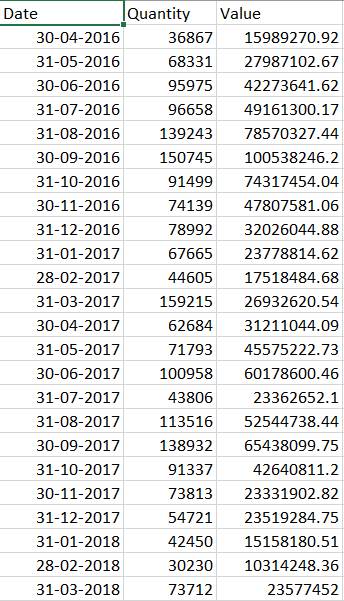
Customer Lifetime Value(CLV):

1. Customer Lifetime Value is an individual score which is given to every year and based on the performance on how they did in that year.
2. CLV is calculated using recency, frequency and monetary of the dealer.
3. To calculate CLV we must normalize the Recency, Frequency and Monetary of each column. To get a an accurate value of the CLV we have done 1-NR to correspond the Recency of the company.
4. We have used this formula of 0.105NR+0.258NM+0.637NF
5. The above weightage is decided by companies specifications, hence it will change for different company.

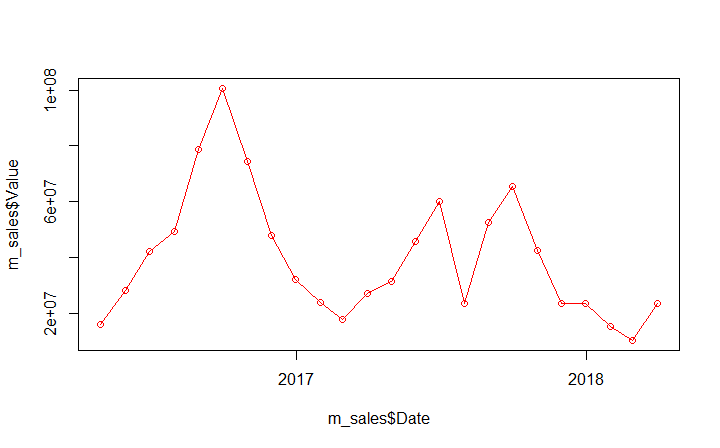


Forcasting:

There are two different datasets for two years. This datasets are having records on daily basis so as first step of pre-processing it is to be aggregated in monthly data. After that, both datasets should be concatenated to get a two years datasets from 2016 April to 2018 April.

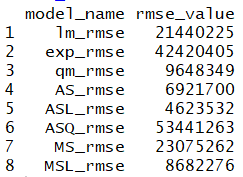


Graph of above Values



Building different Models

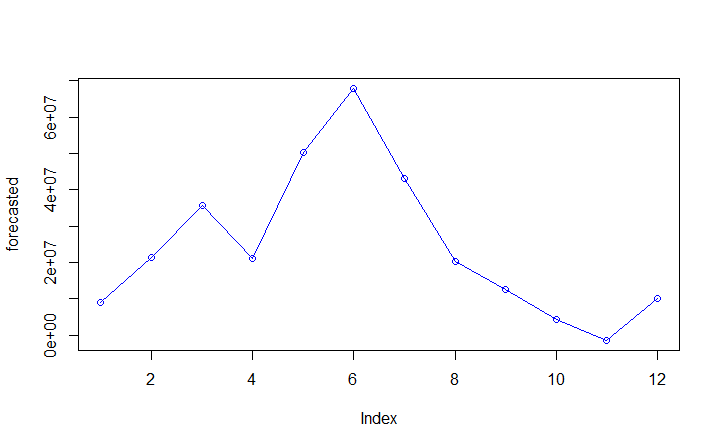
Checking rmse and selecting best model to forecast from below models list.



ASL\_rmse(Additive Seasonality with Linear trend) gives the minimum rmse, hence forecasted its remaining error with arima p = 3,d=0 and q=1 with the help of ACF and PACF plot. At last build model on all historical data and predicted both old data and its error and plotted below showing predicted(green) VS actual(red).



Then created dummy data for further one year and forecasted Sales of year 2018 April to 2019 April in below plot:



ACF and PACF plot showed a strong correlation with 12 month seasonality but because we have total of 24 months arima with p = 12 and q = 12 will give NA values. So I created simple moving average with 12 month as season and plotted the result with above model.

