# Market Analytics and Statistical Modelling

By Saaket Varma
Feb 2019

| | |
|---|---|
| **Market Analytics** | **Introduction:**<br>Marketing analytics is measuring, analyzing, and managing marketing strategy performance to maximize its effectiveness and optimize your return on investment (ROI). Data analytics is an important component of decision making strategy which allow marketers to be more efficient and minimize wasting the marketing budget. Every day there are decisions in the companies to be made. business decision-making process is based on data. With the help of statistical models, we are able to support these decisions. Statistical modeling has an important impact on the performance of the businesses. |
| **Problem Statement** | The sales department of XXXXX company have hired Data Scientists from Springboard to explore their data sets and come up with strategies to provide key business insights in their decision making. |
| **Datasets** | Churn Data, Sales Data, Sales Data-months 2-4, survival data, Default data, News data. |
| **Questions to be addressed** | 1) How can you decide which customers are most valuable for your business? We will model the customer lifetime value using linear regression.<br>2) Predicting if a customer will leave your business, or churn, is important for targeting valuable customers and retaining those who are at risk. We will model customer churn using logistic regression.<br>3) We will model the time to an event using survival analysis. This could be the time until next order or until a person churns.<br>4) How do you analyze CRM data? We will use PCA to condense information to single indices and to solve multicollinearity problems in a regression analysis with many intercorrelated variables. |
| **Methods** | 1) Linear Regression analysis<br>2) Logistic Regression analysis<br>3) PCA component analysis |
| **Significance** | *"Companies making improvements in their measurements and ROI capabilities were more likely to report outgrowing" competitors and a higher level of effectiveness and efficiency in their marketing."* |

# Market Analyses: divided into four parts

- **Modelling The Customer Lifetime Value**
  - **"*salesData*"** customers info for sales for 3 months and the 4th month in two cols.
  - Steps:
    - Perform EDA: correlation, regression to predict sales for the 5th month.
    - Multiple regression analysis and find the best model to explain and predict future sales.
- **Churn Prevention in Marketing**
  - ***"defaultData"*** contains information on customers who are going to default on the loans.
  - Steps:
    - Logistic regression model, model correction and then predict if the customers will churn.
    - Out of sample validation, Cross validation to fine tune the model.
- **Survival Analysis**
  - ***"survivalDataExercise"*** customer info about 'daysSinceFirstPurch' and if they 'boughtAgain' along with sex, shopping cart value, vouchers used and returns.
  - Steps:
    - EDA: distribution of 'daysSinceFirstPurch' against boughtAgain',
    - Survival Analysis using Kaplan-Meier fitting from liflineline library.
    - Inclusion of a covariate, 'voucher'-> multiple variates analysis with CoxPH model
    - Model assumptions and Validating the model
- **CRM data analysis through PCA and Linear Regression.**
  - ***"newsData"*** consists analysis of text and contains variables explaining the text.
  - Steps:
    - correlation, standardization, multiple regression with variables vs PCA.
  - Comparison of two models.

**Data Sets and Analyses**

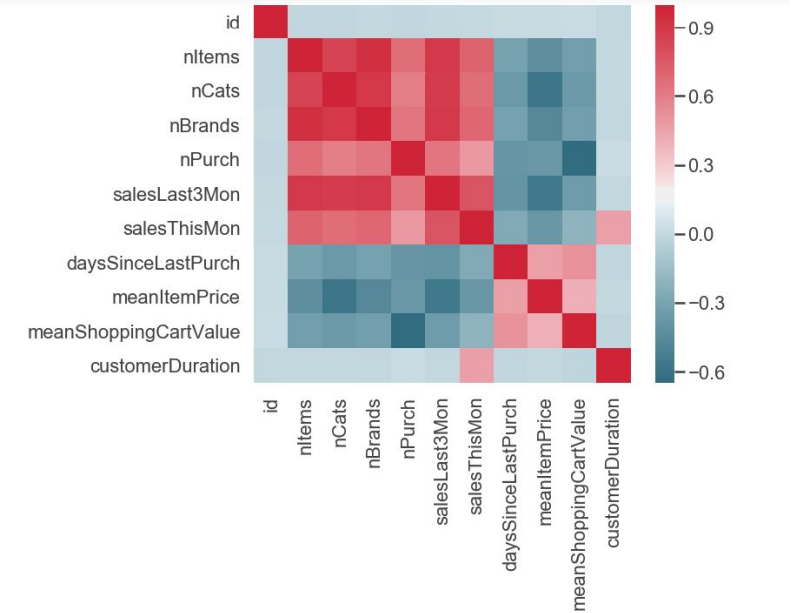
The data sets are clean with no missing values. These dataset were taken from course in R in datacamp.

- source: https://campus.datacamp.com/courses/marketing-analytics-in-r-statistical-modeling/
- Sales Data, Sales Data-months 2-4, (predicting sales using linear regression model)
- Churn Data ( predicting when the customer will churn using logistic regression model)
- survival data, Default data, (Survival analysis using KM and Cox PH models)
- Newsdata. First CLV and Second CLV data sets. (CRM data analysis by PCA and linear regression analysis)

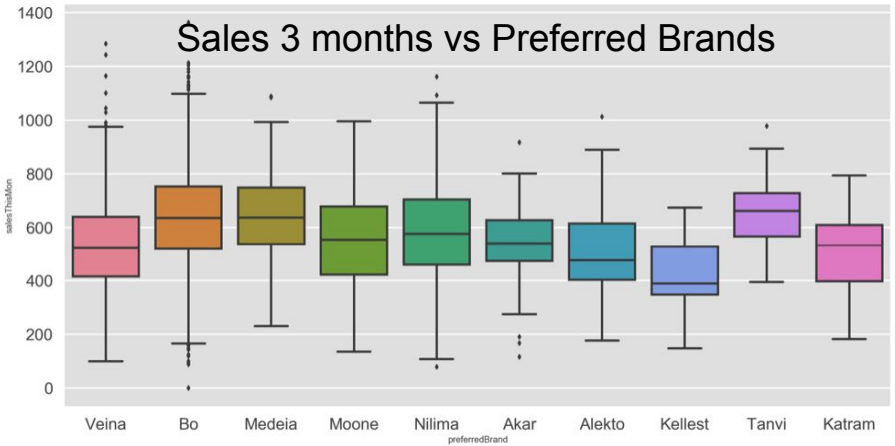
# (I) Modelling the Customer Lifetime Value with Linear Regression

EDA

Correlation matrix between variables



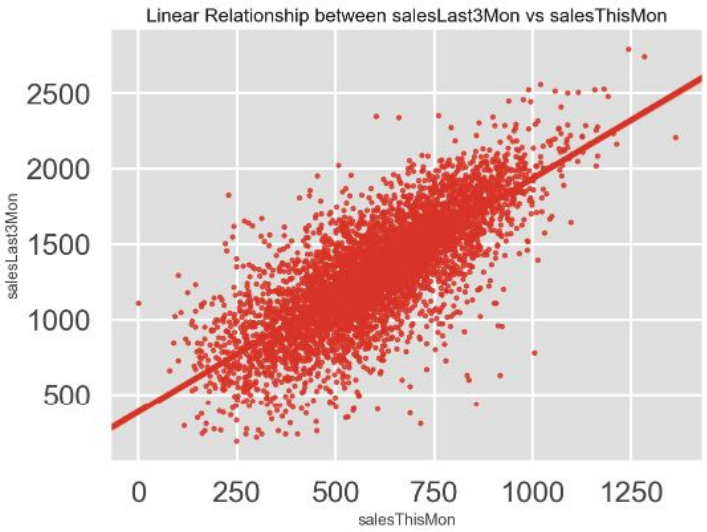





We see a strong correlation between salesLast3Mon & salesThisMon.

We will choose salesLast3Mon as explanatory variable for salesThsMon as a target variable.

# Can sales3month explain salesthismonth?




OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | salesThisMon | R-squared: | 0.593 |
| Model: | OLS | Adj. R-squared: | 0.593 |
| Method: | Least Squares | F-statistic: | 7465. |
| Date: | Mon, 11 Mar 2019 | Prob (F-statistic): | 0.00 |
| Time: | 03:25:19 | Log-Likelihood: | -31680. |
| No. Observations: | 5122 | AIC: | 6.336e+04 |
| Df Residuals: | 5120 | BIC: | 6.338e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 99.6905 | 6.084 | 16.386 | 0.000 | 87.763 | 111.618 |
| salesLast3Mon | 0.3827 | 0.004 | 86.401 | 0.000 | 0.374 | 0.391 |

| | | | |
|---|---|---|---|
| Omnibus: | 174.957 | Durbin-Watson: | 1.956 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 445.273 |
| Skew: | -0.133 | Prob(JB): | 2.04e-97 |
| Kurtosis: | 4.420 | Cond. No. | 5.09e+03 |

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.09e+03. This might indicate that there are strong multicollinearity or other numerical problems.

- The sales3month explain 59% of the change in sales this month.
- We may want add more explanatory variables to improve the model. There might be a problem with multicollinearity.
- We will take care of improving the model and collinearity problem.

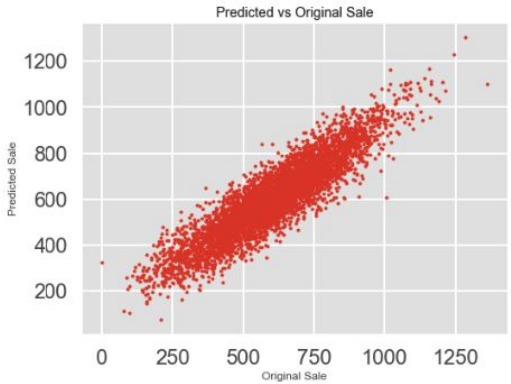
# Summary of two Multiple Regression Models

Multiple regression with all the variables:
mean vif = 12

```
sion Results
==========================================
  R-squared:                       0.825
  Adj. R-squared:                  0.824
  F-statistic:                     665.6
  Prob (F-statistic):               0.00
  Log-Likelihood:                -29521.
  AIC:                         5.912e+04
  BIC:                         5.936e+04
```

Multiple regression after dropping two variables:
mean vif = 5

```
ion Results
==========================================
  R-squared:                       0.824
  Adj. R-squared:                  0.823
  F-statistic:                     914.9
  Prob (F-statistic):               0.00
  Log-Likelihood:                -29540.
  AIC:                         5.913e+04
  BIC:                         5.931e+04
```



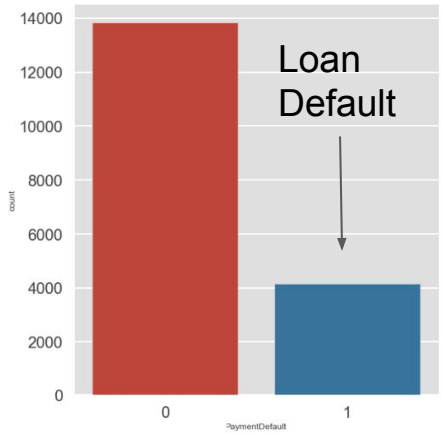

Future prediction of sales for the 5th month = 625 units.

**Multiple Regression to Improve the Model and Prediction of Sales.**

- Improved the regression score from 0.59 to 0.82 by including more variables.
- Multicollinearity problem when using all the variables, vif > 10.
- We took care of the multicollinearity by:
  - vif analysis and removal of variables with vif > 10
    - “preferredBrand” & “nBrands” had vif > 10 so we removed these two variables from the model.
    - Model performance improved as can be seen by similar regression scores & increase in F statistics.
    - Avg. vif dropped from 12 to 5.2 which is acceptable.
- We used this model for making future prediction of sales.
  - We found the mean of 5th month sale = 625.04 units.

# (II) Churn Analysis of Bank Customers who Default on Loans

## EDA




- ~ 1/3 or the bank customers default on payments.
- 4150 pay loans
- 1385 don't pay loans
- Dependent variable is not balanced

## Logistic regression model with all variables

```
Pseudo R-squared:  0.114
AIC:               17276.6520
BIC:               17456.0090
Log-Likelihood:    -8615.3
```

Problems with this model:

- Pseudo R2 score is low
- AIC is score is very high

In sample predictions:

- Accuracy score = 0.77
- Confusion matrix

```
[[13847     3]
 [ 4150     0]]
```

- **We were able to minimize the no. of variables and still achieved similar scores.**
- **In sample and out of sample predictions and validations yielded similar scores suggesting no overfitting.**

## Model Selection using RFE

Selected top 9 variables: 'sex', 'education', 'marriage', 'pay1', 'pay2', 'pay3', 'pay4', 'pay5', 'pay6'

- Pseudo R2 score is low
- AIC is score is very high

When making predictions:

In samples:

- Accuracy score = 0.8
- Confusion matrix

```
[[13472   378]
 [ 3170   980]]
```

Out of sample prediction

- Accuracy score = 0.8
- Confusion matrix

```
[[4032  136]
 [ 939  293]]
```

Cross validations: 5 fold

- Accuracy scores = 0.8

# Conclusion from Churn Analysis

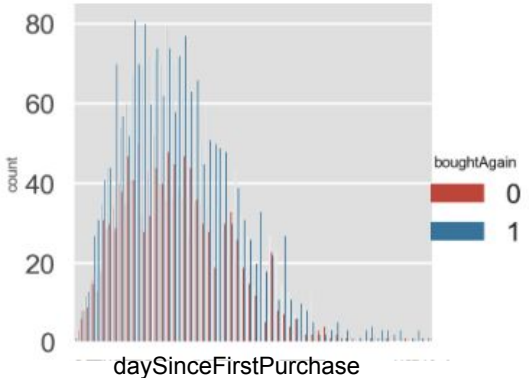
- We analyzed the data for loan payment default.
- Implemented a logistic regression model to predict the probabilities of customers which will churn.
- We fine tuned our model with RFE, minimizing the number of variables and yes not compromising on the accuracy scores.
- Validated our model by in sample and out sample predictions and cross validations.

# (III) Survival Analysis in Customer Relationship Management

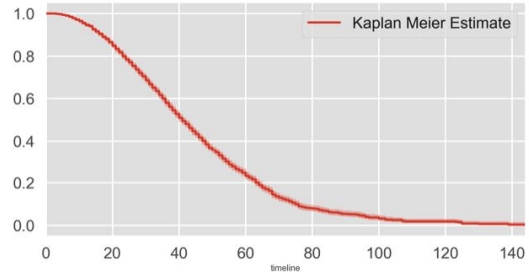
| | daysSinceFirstPurch | shoppingCartValue | gender | voucher | returned | boughtAgain |
|---|---|---|---|---|---|---|
| 0 | 37 | 33.44 | male | 0 | 0 | 0 |
| 1 | 63 | 31.71 | male | 1 | 0 | 1 |
| 2 | 48 | 27.31 | female | 0 | 0 | 0 |
| 3 | 17 | 41.07 | male | 0 | 0 | 1 |
| 4 | 53 | 65.56 | female | 0 | 0 | 0 |

T =tenure

C =churn

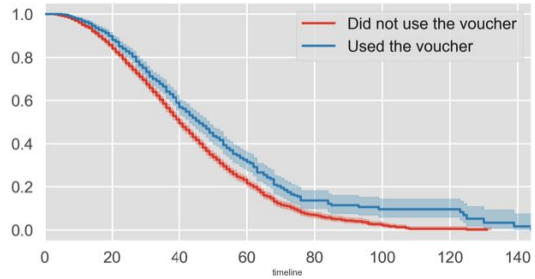



- There are more customers in the data who bought a second time.
- The differences between the distributions are not very large.




No covariate

| n | events | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|
| 5122 | 3199 | 41 | 40 | 42 |



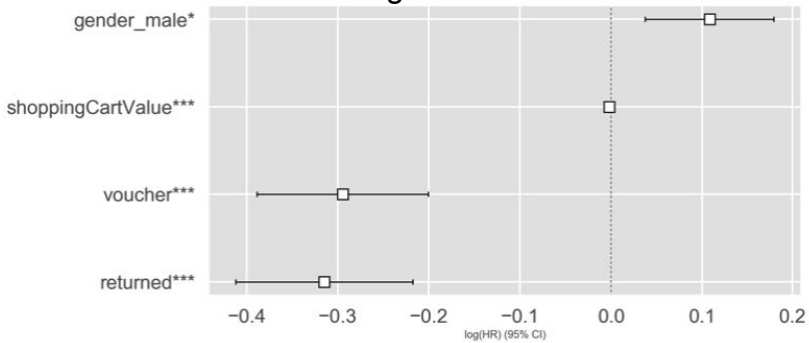

Covariate: Voucher

Customers using a voucher seem to take longer to place their second order. They maybe waiting for another voucher?

# CoxPH Model with Constant Covariates

| | coef | exp(coef) | se(coef) | z | p | log(p) | lower 0.95 | upper 0.95 |
|---|---|---|---|---|---|---|---|---|
| shoppingCartValue | -0.002164 | 0.997839 | 0.000284 | -7.618988 | 2.556723e-14 | -31.297465 | -0.002720 | -0.001607 |
| voucher | -0.294614 | 0.744819 | 0.047969 | -6.141788 | 8.159788e-10 | -20.926633 | -0.388631 | -0.200597 |
| returned | -0.314829 | 0.729914 | 0.049470 | -6.364039 | 1.965164e-10 | -22.350275 | -0.411788 | -0.217869 |
| gender_male | 0.108263 | 1.114341 | 0.036323 | 2.980558 | 2.877233e-03 | -5.850926 | 0.037071 | 0.1794[illegible] |

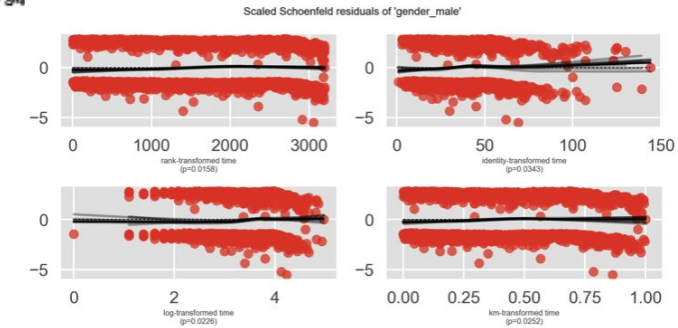
Plotting Hazard ratios




- Shopping cart value increase of 1 dollar decreases the hazard to buy again by a factor of only slightly below 1 - but the coefficient is significant, as are all coefficients.
- For customers who used a voucher, the hazard is 0.74 times lower, and for customers who returned any of the items, the hazard is 0.73 times lower.
- Being a man compared to a woman increases the hazard of buying again by the factor 1.11.

Testing model assumptions by:

- CoxPHFitter.check_assumptions
- Proportional hazard test




Variable 'gender_male' failed the non-proportional test, p=0.0158.

Validation of the model:
R2 score mean = 0.56
Std = 0.022.
Consistent values.

# Conclusions from Survival Analysis

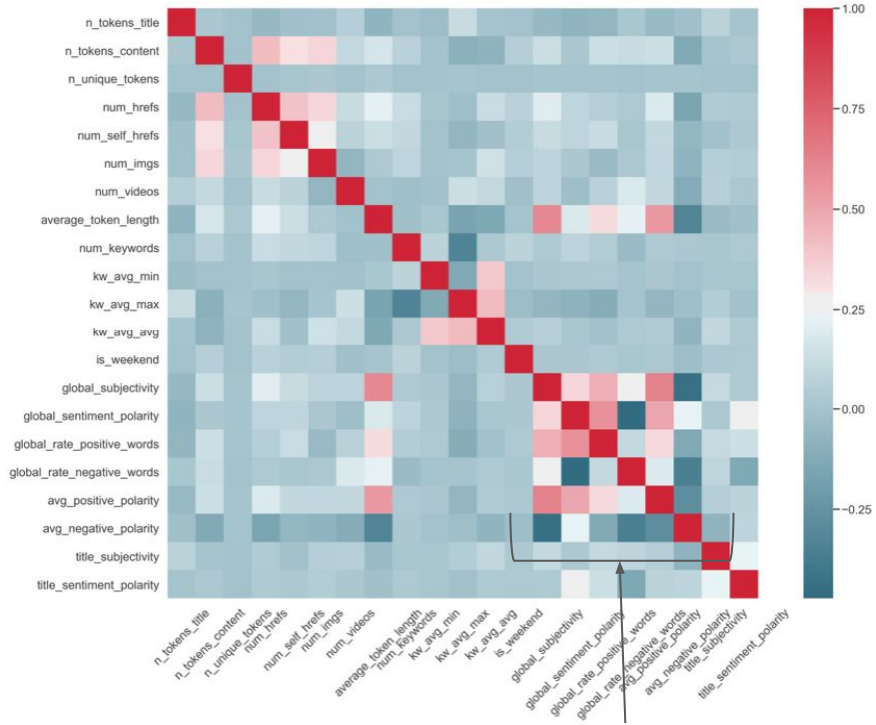
- We performed survival analysis using Kaplan-Meier fitting
- The analysis was performed without and with a categorical covariate, "voucher"
- We also included more variables and performed CoxPH fitting to analyze the data
- We tested the model assumptions and showed thathe model was valid to predict from new data.

Analysis of the data showed the following:

- Customers using a voucher seem to take longer to place their second order.
- Inclusion of other variables showed that a shopper's repurchasing behavior is dependent on the shopping cart value, voucher and customers who returned an item. Sex was not  a significant factor affecting  the purchasing behavior of the customer
- Based on our model the predicted median time until the second order from a female shopper with a voucher is 44 days

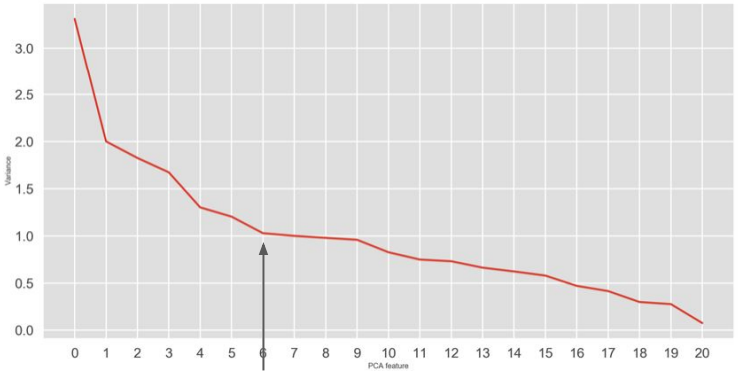
# (IV) CRM data analysis through PCA

EDA Correlation matrix between Variables




Intercorrelated variables in the bottom right area of the plot

Scree plot




- We see an elbow at component 6
- Kaiser-Guttmann rule to drop variance below 1, susggests 8 components.
- Limit our analysis is to top 6 components. (Dimension reduction)

# Which variables form the 6 PCA components?

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| n_tokens_title | -0.053633 | 0.101769 | -0.009280 | 0.099753 | -0.204011 | -0.276678 |
| n_tokens_content | 0.226525 | 0.171174 | 0.380570 | -0.122321 | -0.145989 | -0.017639 |
| n_unique_tokens | 0.004574 | 0.001245 | 0.001103 | -0.009639 | -0.006617 | 0.060556 |
| num_hrefs | 0.258567 | 0.161343 | 0.421143 | 0.034536 | -0.074802 | 0.116485 |
| num_self_hrefs | 0.195081 | 0.073520 | 0.393133 | -0.056833 | -0.122793 | 0.081787 |
| num_imgs | 0.142360 | 0.150678 | 0.431450 | 0.064258 | -0.036204 | 0.079932 |
| num_videos | 0.085118 | 0.195465 | -0.042102 | 0.190159 | -0.163364 | -0.141730 |
| average_token_length | 0.388725 | 0.023326 | -0.194204 | -0.184340 | 0.008825 | 0.142879 |
| num_keywords | 0.074826 | -0.111427 | 0.248698 | -0.143142 | 0.422033 | -0.298207 |
| kw_avg_min | 0.029892 | -0.006099 | 0.047198 | 0.250915 | 0.654126 | 0.069040 |
| kw_avg_max | -0.097255 | 0.207966 | -0.097657 | 0.498130 | -0.344318 | 0.257509 |
| kw_avg_avg | 0.017613 | 0.154115 | 0.057807 | 0.611540 | 0.311769 | 0.174936 |
| is_weekend | 0.046174 | 0.005112 | 0.118472 | 0.021583 | 0.100165 | -0.154984 |
| → global_subjectivity | 0.447540 | 0.006127 | -0.228786 | 0.037198 | 0.030634 | 0.031374 |
| global_sentiment_polarity | 0.249816 | -0.552329 | 0.032066 | 0.184851 | -0.107411 | 0.130848 |
| global_rate_positive_words | 0.333811 | -0.253117 | -0.139516 | 0.077075 | -0.035980 | -0.086754 |
| global_rate_negative_words | 0.147449 | 0.465199 | -0.229438 | -0.114529 | 0.100444 | -0.206262 |
| → avg_positive_polarity | 0.419330 | -0.089454 | -0.172568 | 0.056782 | -0.017682 | 0.094085 |
| avg_negative_polarity | -0.249621 | -0.369361 | 0.200218 | 0.039643 | -0.077033 | 0.065355 |
| title_subjectivity | 0.072301 | 0.027106 | -0.009466 | 0.267040 | -0.073111 | -0.613880 |
| title_sentiment_polarity | 0.067951 | -0.239211 | 0.105357 | 0.244040 | -0.145055 | -0.423730 |

- **PC1** reflects “Subjectivity” (high global_subjectivity and avg_positive_polarity, negative loading on avg_negative_polarity).
- **PC2** contains “Positivity” (high global_sentiment_polarity, low global_rate_negative_words; even negative words are not very negative as you can see from the positive loading on avg_negative_polarity).
- We see that the same group of intercorrelated variables from the corrplot, but they split into two components.

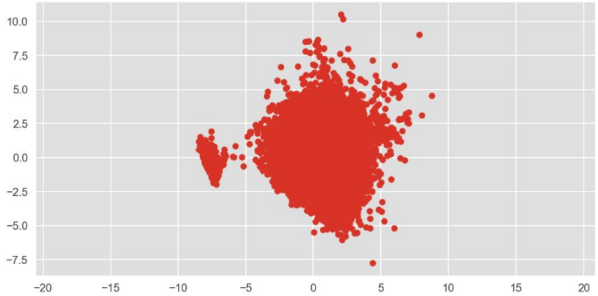
# Comparison of Regression Analysis:

## (A) Many Variables Vs (B) PCA components

| | VIF Factor | features |
|---|---|---|
| 0 | 19.413792 | n_tokens_title |
| 1 | 3.335565 | n_tokens_content |
| 2 | 1.026498 | n_unique_tokens |
| 3 | 2.948667 | num_hrefs |
| 4 | 2.170397 | num_self_hrefs |
| 5 | 1.700050 | num_imgs |
| 6 | 1.222423 | num_videos |
| 7 | 45.637402 | average_token_length |
| 8 | 14.233546 | num_keywords |
| 9 | 1.727662 | kw_avg_min |
| 10 | 7.427232 | kw_avg_max |
| 11 | 12.343251 | kw_avg_avg |
| 12 | 1.170368 | is_weekend |
| 13 | 40.024822 | global_subjectivity |
| 14 | 17.093201 | global_sentiment_polarity |
| 15 | 18.139153 | global_rate_positive_words |
| 16 | 10.423926 | global_rate_negative_words |
| 17 | 43.208810 | avg_positive_polarity |
| 18 | 9.450196 | avg_negative_polarity |
| 19 | 1.975245 | title_subjectivity |
| 20 | 1.215339 | title_sentiment_polarity |

- $R^2$ score = 0.08
- 9 features have more than vif > 10 showing multicollinearity.
- Vif mean = 12.18
- > 10 the model is not suitable and suggests high multicollinearity

Pearson correlation between PCA1 & PCA2




There is no correlation between PCA1 & PCA2

| | |
|---|---|
| R-squared: | 0.051 |
| Adj. R-squared: | 0.051 |
| F-statistic: | 353.2 |
| Prob (F-statistic): | 0.00 |
| Log-Likelihood: | -52363. |
| AIC: | 1.047e+05 |
| BIC: | 1.048e+05 |

- Linear regression analysis with PCA components, R squared decreased only from 8% to 5% even though the number of variables decreased from 21 to 6.
- There was no correlation between PCA components.

# Conclusions from PCA analysis

- CRM data analysis by reducing the number of variables without reducing too much information.
- We were able to interpret selected components and match it with the fratures/ variables.
- Scree plot and Kaiser-Guttmann method to minimize the no. of PCA components used for analysis.

Further linear regression analysis was performed using these condensed data or PCA components and compared it to the analysis with all the variables selected.

- We found that 6 components although are enough to predict customer activity and brand awareness, the model is not robust. We may want to add more variables for our analysis.
- Using these six components decreased the explained variance, but solves the multicollinearity problem.

# Summary

We performed:

- Linear Regression Analysis to understand the customer lifetime value prediction.
- Logistic Regression to predict Churn.
- Implement Survival Analysis to prevent Churn
- PCA to reduce the dimensionality of the data

Future scope in Market Analytics and other methods used are:

- Cluster Analysis to segment of customers
- Implement factor analysis for customer satisfaction surveys.