Dated: 01/02/2019

Project Ideas: Capstone Project-2

Objective: Capstone Project 2	**To implement machine learning tools in solving real life problems.**
Project Idea (I)	**Market Analytics / Statistical Modeling:** Companies have to make important decisions every day. Statistical models help companies make hypothesis-driven business decisions based on the market data. In this project for the capstone project, I propose to explore the sales data and help them in deciding which customers are important for their business and growth. **Data Sets:** Churn Data, Sales Data, Sales Data-months 2-4, survival data, Default data, News data. **Questions:** ● Which customers are valuable to your business? ● Which customers will leave your business? ● Survival Analysis. ● CRM data Analysis.
Project Idea (II)	**HR Analytics/ People Analytics / Workforce Analytics:** For companies it is becoming more important every day, to hire the best talent and maintain a low turnover rate. At the same time, they want their existing employees to keep their productivity and engagement levels high and accidents a lowest. In this project, I propose to analyze several data sets to address these issues through statistical **Data Set:** a) HR data (2940, 4) b) Accident data (302,3), c) Survey data, d) Performance Data. **Questions:** ● When and where is the highest accident rate? ● Explaining the reasons for increase in accidents? ● Regression to analyze drivers. ● What is driving low employee engagement?
Project Idea (III)	**Fraud Detection in Python:** An organization loses anywhere from 5% to 20% of their revenue to fraudulent activities. Data Scientists at Springboard have been asked by several companies to propose methods to detect fraud. They have previous data we can use and they also want help in future to detect fraudulent activities. **Data Sets:** a) Data Set from Kaggle b) From Datacamp c) Another one **Questions:** ● How to find fraud? ● How to detect fraudulent behavior similar to past ones through supervised learning. ● Learning methods to discover new types of fraud activities through unsupervised learning. ● What is driving low employee engagement? ● How to finding fraud in the text data through NLP?

Market Analytics and Statistical Modeling

Market Analytics	**Introduction**: **Marketing analytics** is measuring, analyzing, and managing **marketing** 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. First CLV and Second CLV data sets.
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. 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. 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. 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 2) Logistic Regression analysis 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.”* *KPI metrics: Sales Revenue, CPC, Online Marketing (ROI), Form Conversion Rates, Social media Search,*

Market Analytics and Statistical Modelling

Introduction

Marketing analytics is measuring, analyzing, and managing marketing 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

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 uisng KM nad Cox PH models)
- News data. First CLV and Second CLV data sets. (CRM data analysis by PCA and linear regression analysis)

Questions to be addressed

- How can you decide which customers are most valuable for your business?
- We will model the customer lifetime value using linear regression. 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.
- We will model the time to an event using survival analysis. This could be the time until next order or until a person churns.
- 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 Used for Data Analysis

- Linear Regression analysis followed by evaluation of the model and prediction of sales.
- Logistic Regression analysis to predict customer churn.
- Survival Analysis through Kaplan-Meier and Cox PH methods.
- PCA component analysis of the CRM data followed by linear regression, model selection and fine tuning the model.

Significance

“Companies making improvements in their measurements and ROI capabilities are more likely to report outgrowing” competitors and a higher level of effectiveness and efficiency in their marketing.” Market analytics and statistical modelling provides exact quantitative information. Beyond the obvious sales and lead generation applications, marketing analytics can offer profound insights into customer preferences and trends.source:https://www.wordstream.com/marketing-analytics

Market Analysis is divided into five subsections:

(I) Modelling & Predicting The Customer Lifetime Value

- **“*salesData*”** customers info for sales for 3 months and the 4th month in two cols.
- Steps:
    - Perform EDA: correlation, regression analysis to predict sales for the 5th month.
    - Multiple regression analysis and
    - Find the best model to explain and predict future sales.
- Conclusions

- **(II) 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.
    - Conclusions

- **(III) Survival Analysis in Customer Relationship Management**
    - ***“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’ and
        - perform multiple variates analysis using CoxPH model.
        - Model assumptions and Validating the model to make sure that there is no overfitting.
    - Conclusions

- **(IV) 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.
    - Conclusions

- **(V) Project Summary of Market Analytics and Statistical Modeling for Capstone Project-2**

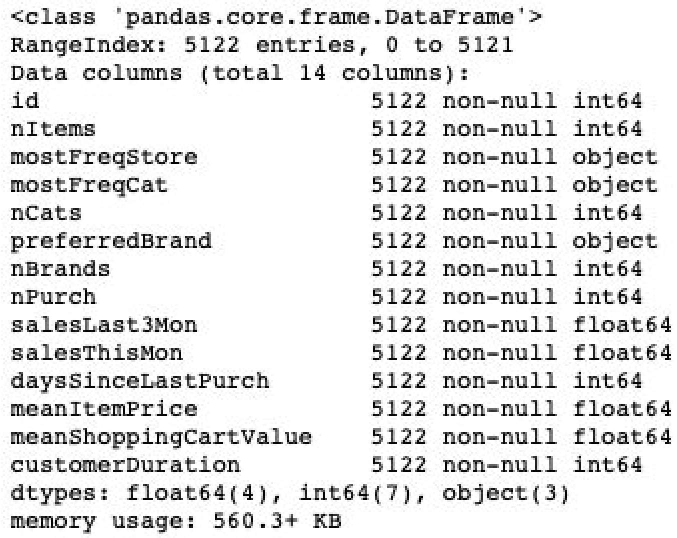
(I) Modelling & Predicting The Customer Lifetime Value

(I) Modeling and Predicting The Customer Lifetime Value

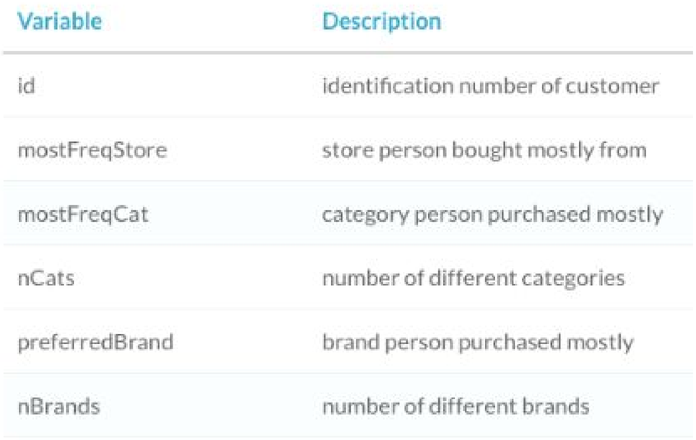
Exploratory Data Analysis:

We will explore the data through heatmaps, pair plots and correlation plots. We will also compare the distribution of target variables across categorical variables by visualization with boxplots.

SalesData info:



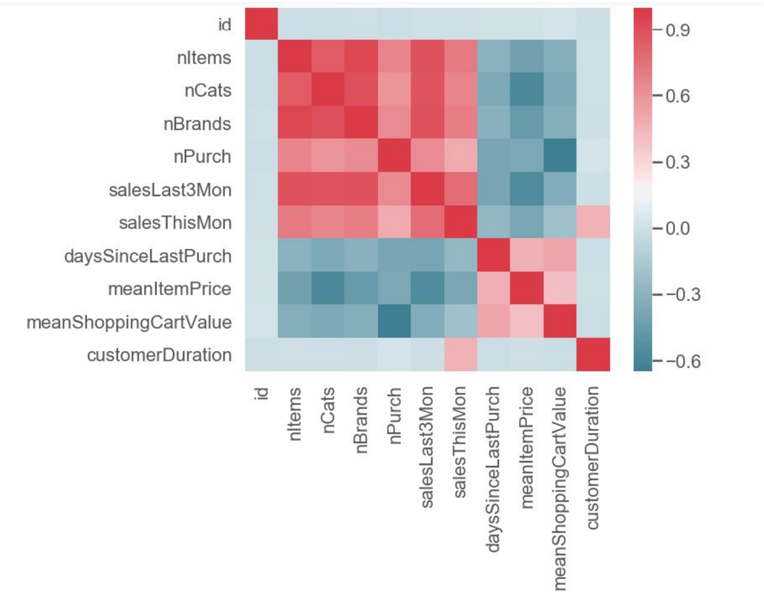
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5122 entries, 0 to 5121
Data columns (total 14 columns):
id                        5122 non-null int64
nItems                    5122 non-null int64
mostFreqStore             5122 non-null object
mostFreqCat               5122 non-null object
nCats                     5122 non-null int64
preferredBrand            5122 non-null object
nBrands                   5122 non-null int64
nPurch                    5122 non-null int64
salesLast3Mon             5122 non-null float64
salesThisMon              5122 non-null float64
daysSinceLastPurch        5122 non-null int64
meanItemPrice             5122 non-null float64
meanShoppingCartValue     5122 non-null float64
customerDuration          5122 non-null int64
dtypes: float64(4), int64(7), object(3)
memory usage: 560.3+ KB
```



Variable	Description
id	identification number of customer
mostFreqStore	store person bought mostly from
mostFreqCat	category person purchased mostly
nCats	number of different categories
preferredBrand	brand person purchased mostly
nBrands	number of different brands

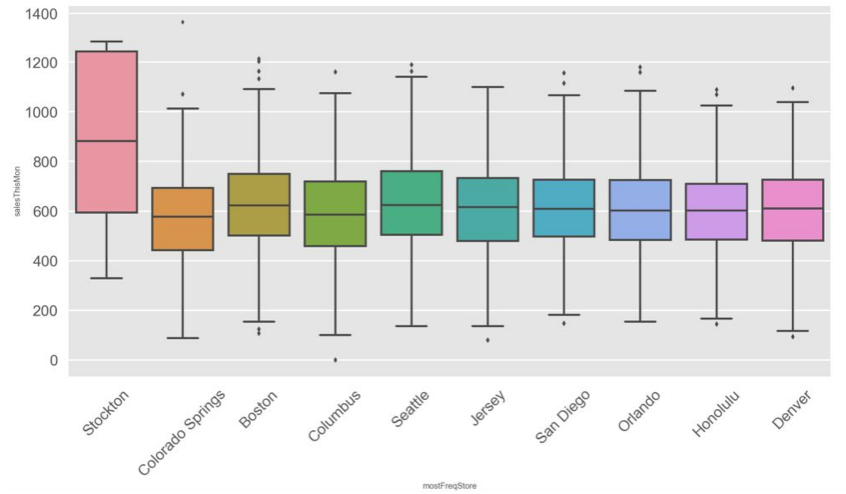
Correlation between the variables

We observe a correlation between ‘salesLast3Mon’ and ‘salesThismon’

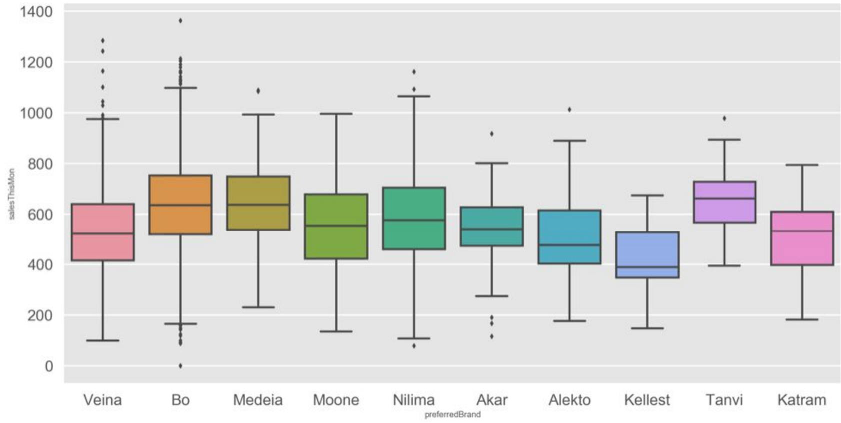


Further exploration and observing the correlation of sales over 3 months and categorical variables.

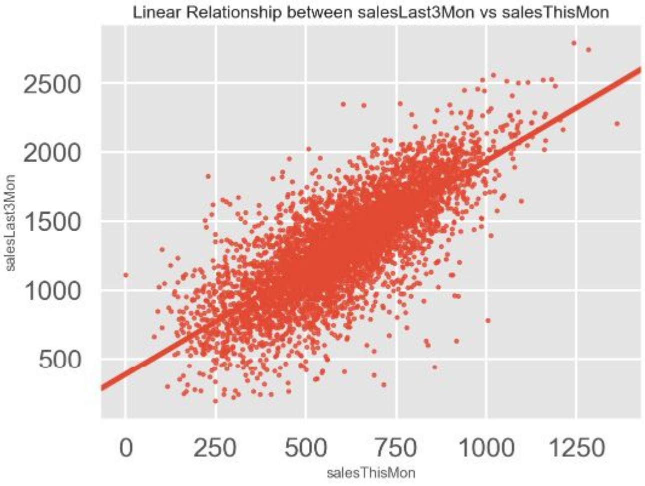
Sales for last 3 months vs Stores (“mosFrequentStore”)



Sales for last 3 months vs Preferred Brands

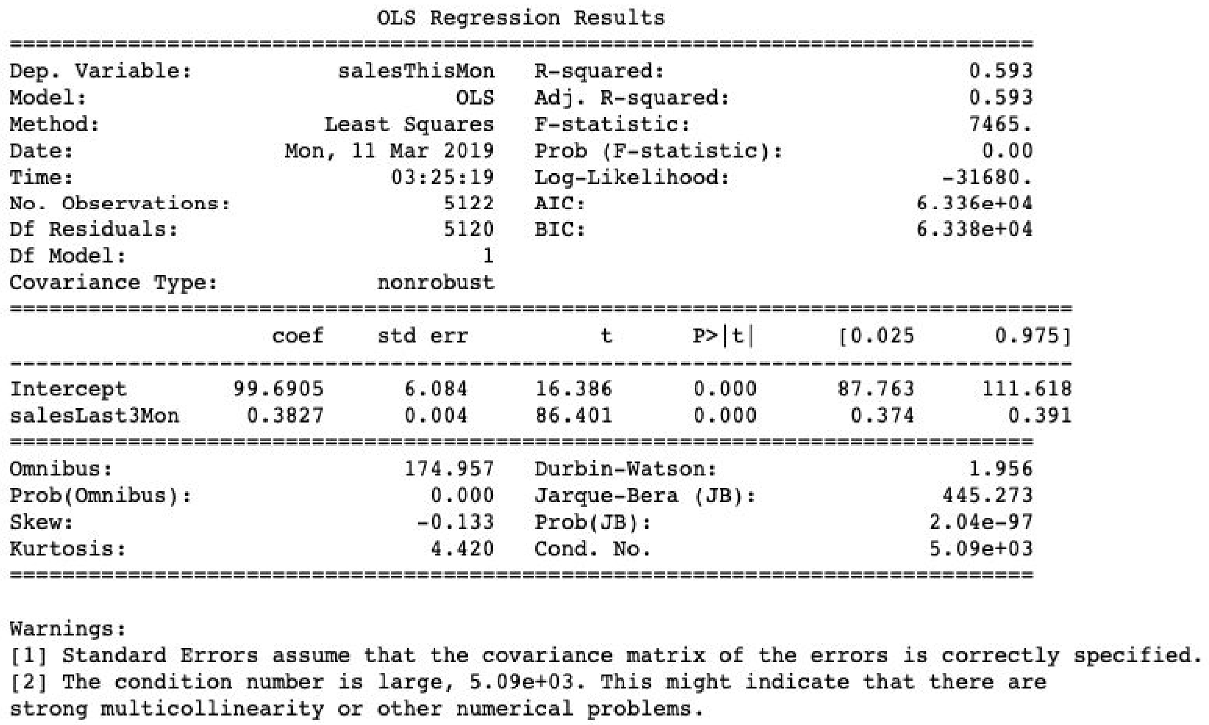


We saw that the sales in the last three months are strongly positively correlated with the sales in this month. Hence we will start off including that as an explanatory variable in a linear regression. Lets see the correlation in a scatter plot.



We observe a high correlation, hence for creating a model which can predict future sales, we will choose ‘salesLast3Mon’ as explanatory variable for ‘salesThisMon’ as a target variable.

Linear Regression Analysis



```
                            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.
```

Observations from the linear regression model:

- 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 also take care of improving the model and collinearity problem.

Multiple Regression Models

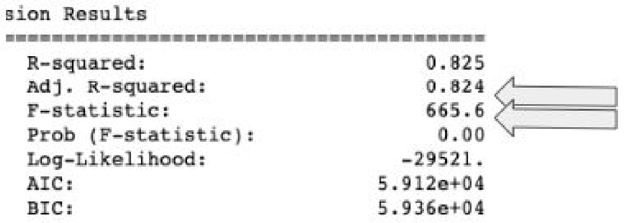
A multiple regression model, refers to regression models with only one dependent/outcome variable and many independent/predictor variables. source
https://www.quora.com/What-is-multivariate-regression

First we used all the variables to explain the sales for the current month. Variance inflation factor was higher than 10 so in order to take care of multicollinearity problem we removed 2 variables,

Below is a comparison of the 2 multiple regression models. (PTO)

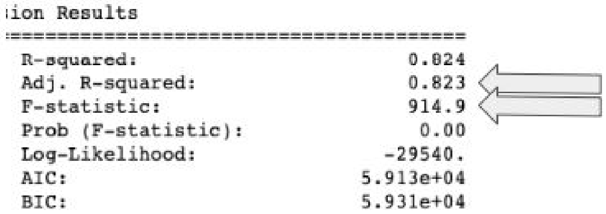
Summary of two Multiple Regression Models

Multiple regression with all the variables:
mean vif = 12

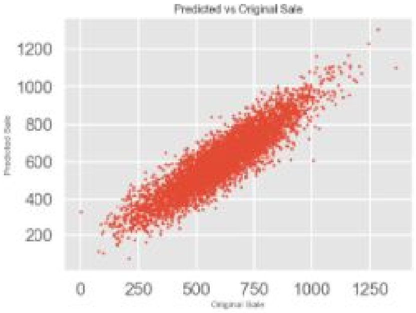


```
sion Results
==========================================
  R-squared:                         0.825
  Adj. R-squared:                    0.824
  F-statistic:                       665.6
  Prob (F-statistic):                 0.00
  Log-Likelihood:                  -29521.
  AIC:                           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.

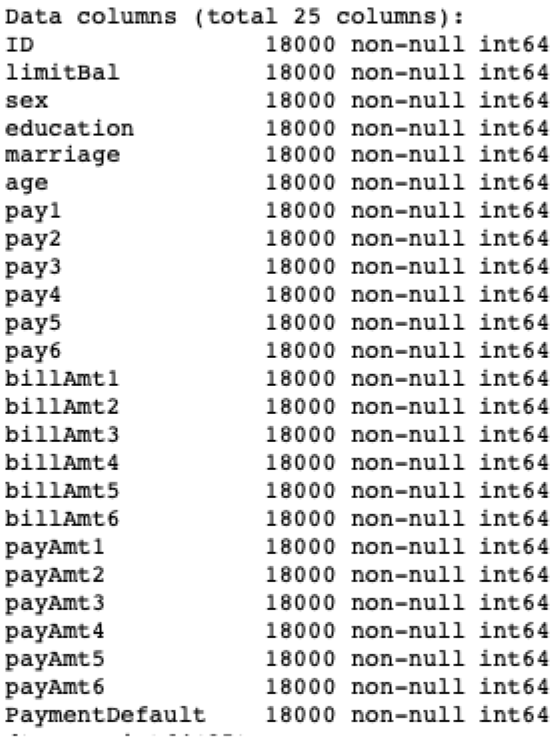
Conclusion of Part I: Customer life-time value (in this case sales)

- 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.

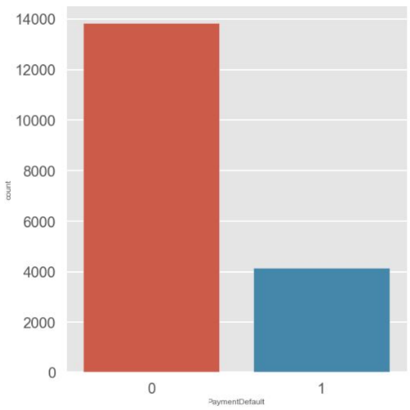
(II) Churn Analysis of Bank Customers who Default in Loans using Logistic Regression

(II) Churn Analysis of Bank Customers who Default in Loans using Logistic Regression

Exploration Data Analysis



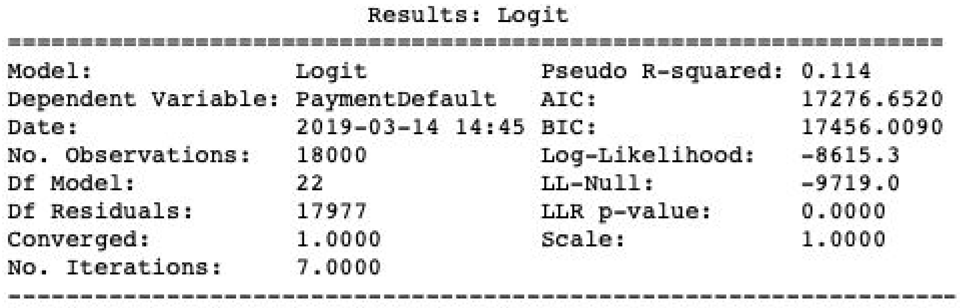
```
Data columns (total 25 columns):
ID                  18000 non-null int64
limitBal            18000 non-null int64
sex                 18000 non-null int64
education           18000 non-null int64
marriage            18000 non-null int64
age                 18000 non-null int64
pay1                18000 non-null int64
pay2                18000 non-null int64
pay3                18000 non-null int64
pay4                18000 non-null int64
pay5                18000 non-null int64
pay6                18000 non-null int64
billAmt1            18000 non-null int64
billAmt2            18000 non-null int64
billAmt3            18000 non-null int64
billAmt4            18000 non-null int64
billAmt5            18000 non-null int64
billAmt6            18000 non-null int64
payAmt1             18000 non-null int64
payAmt2             18000 non-null int64
payAmt3             18000 non-null int64
payAmt4             18000 non-null int64
payAmt5             18000 non-null int64
payAmt6             18000 non-null int64
PaymentDefault      18000 non-null int64
```



We observe that there are:

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

Logistic regression model with all variables to model and predict the probabilities of customers who will default.



```
                              Results: Logit
=================================================================
Model:                Logit             Pseudo R-squared: 0.114
Dependent Variable:   PaymentDefault    AIC:              17276.6520
Date:                 2019-03-14 14:45  BIC:              17456.0090
No. Observations:     18000             Log-Likelihood:   -8615.3
Df Model:             22                LL-Null:          -9719.0
Df Residuals:         17977             LLR p-value:      0.0000
Converged:            1.0000            Scale:            1.0000
No. Iterations:       7.0000
------------------------------------------------------------------
```

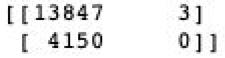
Problems with this model:

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

In sample predictions:

- Accuracy score = 0.77

- Confusion matrix



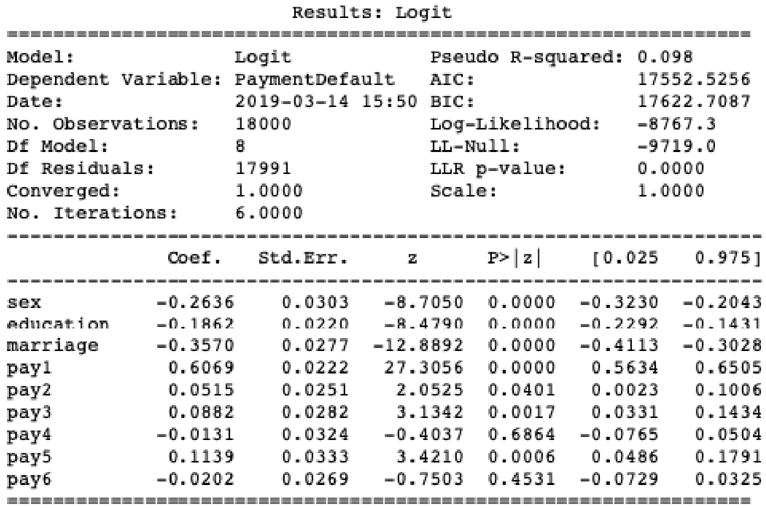
```
[[13847     3]
 [ 4150     0]]
```

The result is telling us that we have 13847+2 correct predictions and 4148+3 incorrect predictions.

Model Selection using RFE

Recursive Feature Elimination (RFE) as its title suggests recursively removes features, builds a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable (or class). Scikit Learn does most of the heavy lifting just import RFE from sklearn.feature_selection and pass any classifier model to the RFE() method with the number of features to select. Using familiar Scikit Learn syntax, the .fit() method must then be called.

Through RFE we selected top 9 variables: 'sex', 'education', 'marriage', 'pay1', 'pay2', 'pay3', 'pay4', 'pay5', 'pay6'.



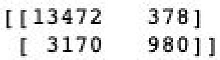
```
                        Results: Logit
===================================================================
Model:              Logit            Pseudo R-squared: 0.098
Dependent Variable: PaymentDefault   AIC:              17552.5256
Date:               2019-03-14 15:50 BIC:              17622.7087
No. Observations:   18000            Log-Likelihood:   -8767.3
Df Model:           8                LL-Null:          -9719.0
Df Residuals:       17991            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
--------------------------------------------------------------------
              Coef.  Std.Err.     z     P>|z|   [0.025   0.975]
--------------------------------------------------------------------
sex          -0.2636   0.0303  -8.7050  0.0000  -0.3230  -0.2043
education    -0.1862   0.0220  -8.4790  0.0000  -0.2292  -0.1431
marriage     -0.3570   0.0277 -12.8892  0.0000  -0.4113  -0.3028
pay1          0.6069   0.0222  27.3056  0.0000   0.5634   0.6505
pay2          0.0515   0.0251   2.0525  0.0401   0.0023   0.1006
pay3          0.0882   0.0282   3.1342  0.0017   0.0331   0.1434
pay4         -0.0131   0.0324  -0.4037  0.6864  -0.0765   0.0504
pay5          0.1139   0.0333   3.4210  0.0006   0.0486   0.1791
pay6         -0.0202   0.0269  -0.7503  0.4531  -0.0729   0.0325
===================================================================
```

Problems with this model:

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

In sample predictions:

- Accuracy score = 0.098
- Confusion matrix



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

The result tells us that we have 13472+980 correct predictions and 3178+980 incorrect predictions.

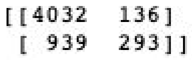
Avoiding overfitting

In order to avoid overfitting we perform out of sample predictions and look at the accuracy scores, followed by 5 fold cross validations to see the distribution of scores.

To perform out of sample validation we perform train test split and train the model on the training set and then make predictions on the test split. Check the accuracy scores and compare it to the earlier models.

Out of sample prediction

- Accuracy score = 0.8
- Confusion matrix



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

The result tells us that we have 4032 +293 correct predictions and 939+136 incorrect predictions.

Cross validations: 5 fold

- Accuracy scores = 0.8

Conclusion of Part II: 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.

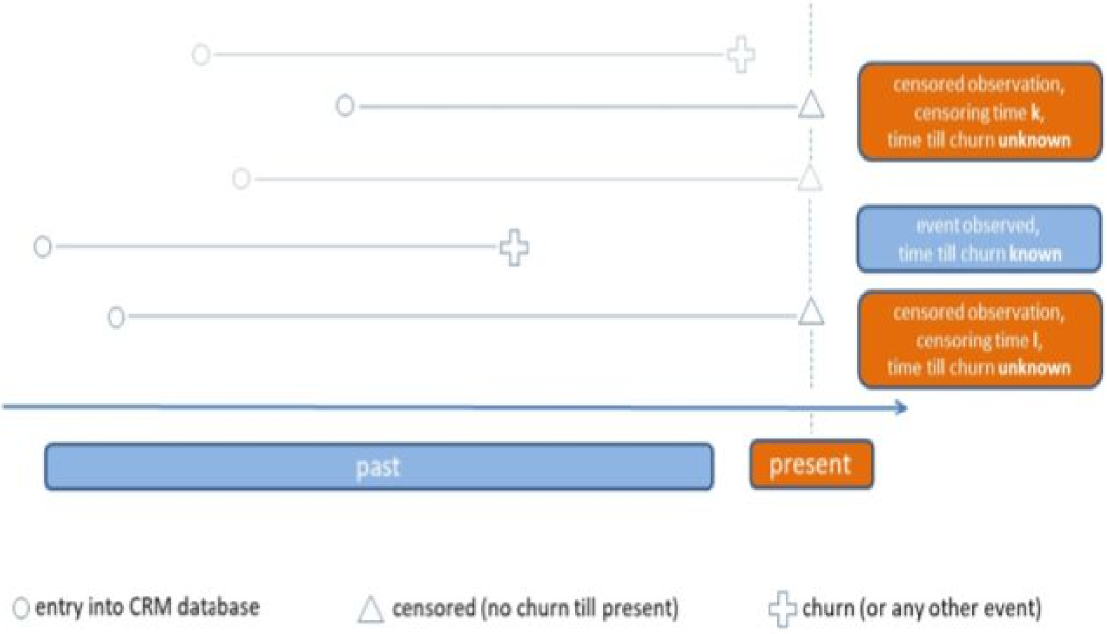
(III) Survival Analysis in Customer Relationship Management

(III) Survival Analysis in Customer Relationship Management

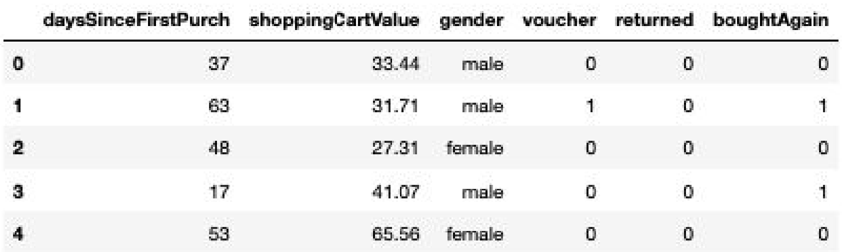
Why survival analysis? We conduct survival analysis to study customer relationship when we have censored data or missing data. Logistic regression cannot handle the censerode data or the missing data.

Advantages of survival analysis:

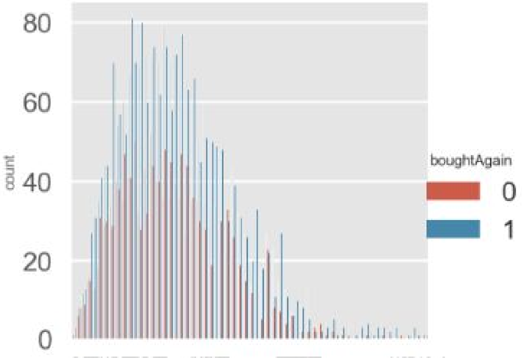
- Allows for model time of an event thus avoids loss of function due to aggregation,
- Allows to model ""when"" an event take place and not just if it will take place,
- No arbitrary set time frame,
- Provides deeper insights into customer relationships.


censored observation, censoring time **k**, time till churn **unknown**
event observed, time till churn **known**
censored observation, censoring time **l**, time till churn **unknown**
past
present
entry into CRM database
censored (no churn till present)
churn (or any other event)

EDA



	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



- The variable boughtAgain takes the value 0 for customers with only one order and 1 for customers who have placed a second order already.
- If a person has ordered a second time, you see the number of days between the first and second order in the variable daysSinceFirstPurch. For customers without a second order, daysSinceFirstPurch contains the time since their first (and most recent) order

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

Preparing the data for Survival Analysis: We need two columns into considerations to build the model.
1) Time under obsetvation: "daysSinceFirstPurch". T= tenure

2) Status at the end of this time: "boughtAgain" This will help us with knowing whether the observation was censored or not. C= churn.

Survival analysis using Kaplan-Meier Fitting - without and with a categorical covariate

Kaplan-Meier analysis allows estimation of survival over time, even when pts drop out or are studied for different lengths of time.

How does it work?

1. For each interval, survival probability is calculated as No. of points surviving / by no. of points at risk.
2. Pts who have died, dropped out, or not reached the time yet are not counted as “at risk.”

Pts who are lost are considered “censored” & are not counted in the denominator.

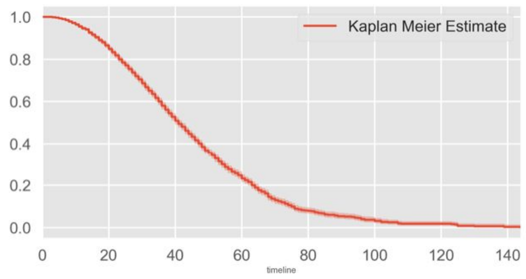
• Probability of surviving to any point is estimated from cumulative probability of surviving each of the preceding time intervals (calculated as the product of preceding probabilities).

• Although the probability calculated at any given interval isn’t very accurate because of the small no. of events, the overall probability of surviving to each point is more accurate.

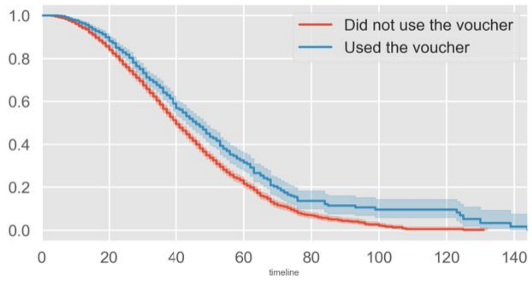
source:

- http://savvastjortjoglou.com/nfl-survival-analysis-kaplan-meier.html
- http://biostat.mc.vanderbilt.edu/wiki/pub/Main/ClinStat/km.lam.pdf

No covariate



Covariate: Voucher



```
n         events  median   0.95LCL      0.95UCL
5122    3199        41        40                  42
```

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

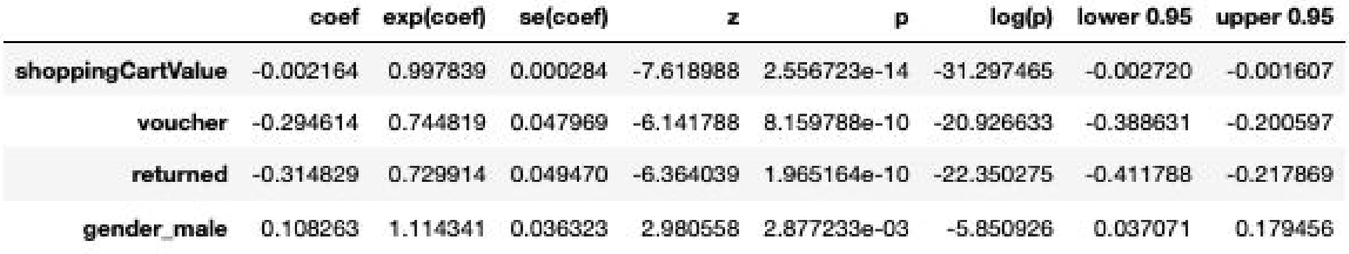
CoxPH Model with Constant Covariates

Why Cox propotional hazard (PH) function in Survival Analysis?

Kaplan-Meier curves and logrank tests are useful only when the predictor variable is categorical (e.g.: treatment A vs treatment B; males vs females). They don’t work easily for quantitative predictors such as gene expression, weight, or age.

An alternative method is the Cox proportional hazards regression analysis, which works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

Note: The coefficients in Cox PH model are interpreted similar to logistic regression model.



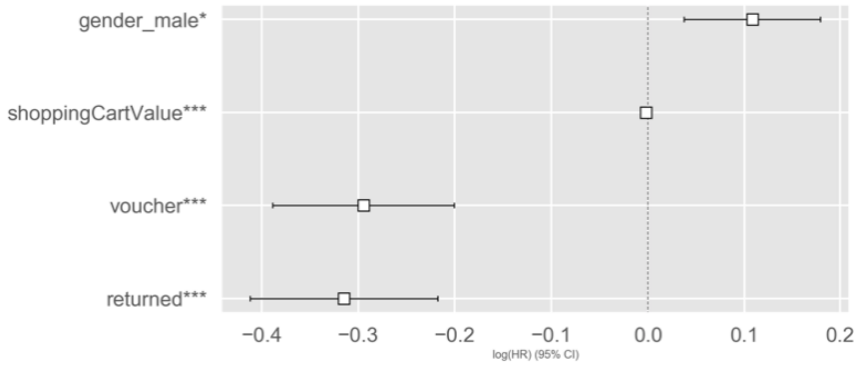
	coef	exp(coef)	se(coef)	z	p	log(p)	lower 0.95	upper 0.95
shoppingCartValue	-0.002164	0.997839	0.000284	-7.618988	2.556723e-14	-31.297465	-0.002720	-0.001607
voucher	-0.294614	0.744819	0.047969	-6.141788	8.159788e-10	-20.926633	-0.388631	-0.200597
returned	-0.314829	0.729914	0.049470	-6.364039	1.965164e-10	-22.350275	-0.411788	-0.217869
gender_male	0.108263	1.114341	0.036323	2.980558	2.877233e-03	-5.850926	0.037071	0.179456

Interpretation of results

Hazard ratios: The hazard ratio can be interpreted as the chance of an event occurring in the treatment arm divided by the chance of the event occurring in the control arm

Hazard ratios differ from relative risks and odds ratios in that RRs and ORs are cumulative over an entire study, using a defined endpoint, while HRs represent instantaneous risk over the study time period, or some subset thereof. Hazard ratios suffer somewhat less from selection bias with respect to the endpoints chosen and can indicate risks that happen before the endpoint.

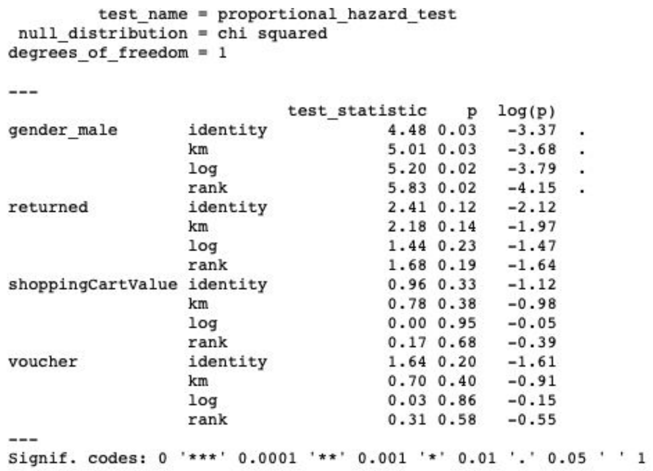
Plotting Hazard ratios of explanatory variables.



- 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.

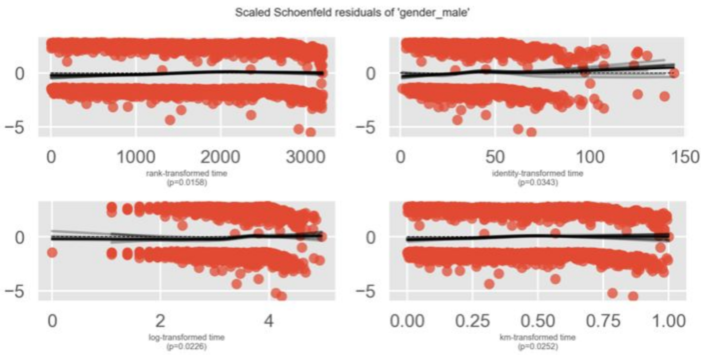
Checking model assumptions

Method 1: We can use the check_assumption_ function to check the validity of the model. source:https://lifelines.readthedocs.io/en/latest/jupyter_notebooks/Proportional%20hazard%20assumption.html#Checking-assumptions-with-check_assumptions



```
           test_name = proportional_hazard_test
   null_distribution = chi squared
  degrees_of_freedom = 1

---
                                    test_statistic    p  log(p)
gender_male       identity                    4.48 0.03   -3.37  .
                  km                          5.01 0.03   -3.68  .
                  log                         5.20 0.02   -3.79  .
                  rank                        5.83 0.02   -4.15  .
returned          identity                    2.41 0.12   -2.12
                  km                          2.18 0.14   -1.97
                  log                         1.44 0.23   -1.47
                  rank                        1.68 0.19   -1.64
shoppingCartValue identity                    0.96 0.33   -1.12
                  km                          0.78 0.38   -0.98
                  log                         0.00 0.95   -0.05
                  rank                        0.17 0.68   -0.39
voucher           identity                    1.64 0.20   -1.61
                  km                          0.70 0.40   -0.91
                  log                         0.03 0.86   -0.15
                  rank                        0.31 0.58   -0.55
---
Signif. codes: 0 '***' 0.0001 '**' 0.001 '*' 0.01 '.' 0.05 ' ' 1
```



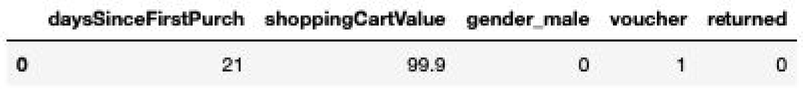
We see that Variable 'gender_male' failed the non-proportional test, p=0.0158.

Validate the model to test for overfitting

The mean scores for the K =10 fold validation = 0.56 with sd = 0.021. which is very tight.

As we can see from the scores the explanatory power of our model is rather low. We could try to collect more explanatory variables and improve the model. This is outside the scope of this analysis.

Making predictions on new customer data



	daysSinceFirstPurch	shoppingCartValue	gender_male	voucher	returned
0	21	99.9	0	1	0

n	events	median	0.95LCL	0.95UCL
5122	3199	47	44	49

We were informed that due to database problems the gender was incorrectly coded: The new customer is actually male. We will change that in our dataframe newCustomer, change the respective variable from femaie to male

n	events	median	0.95LCL	0.95UCL
5122	3199	44	42	47

Conclusion of Part III: 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

(IV) CRM data analysis through PCA

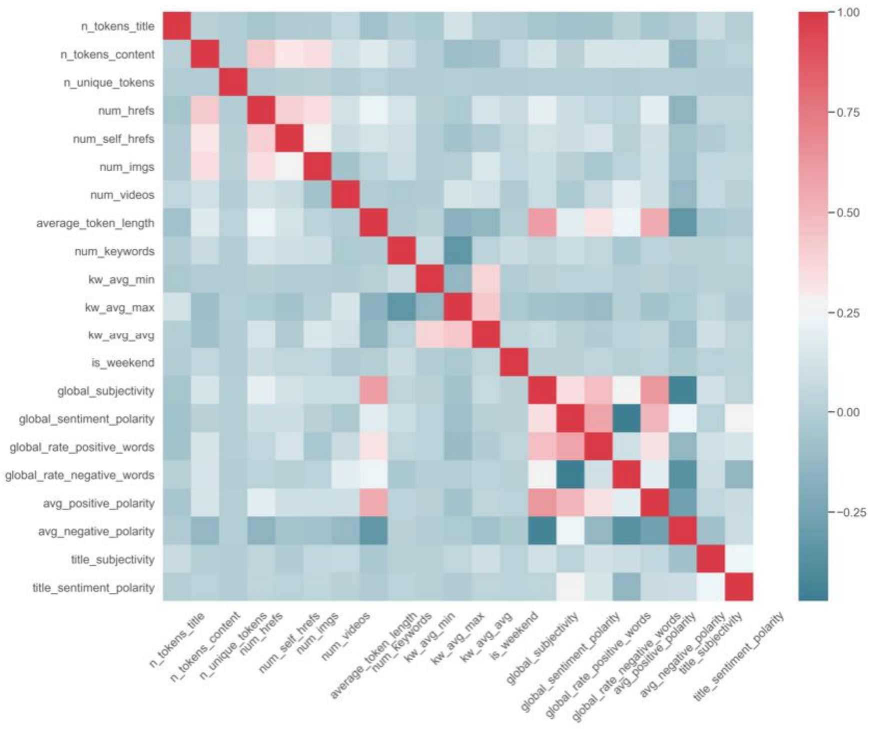
What IS CRM?

Customer relationship management (CRM) is the combination of practices, strategies and technologies that companies use to manage and analyze customer interactions and data throughout the customer lifecycle, with the goal of improving customer service relationships and assisting in customer retention and driving sales growth. CRM systems compile customer data across different channels, or points of contact between the customer and the company, which could include the company's website, telephone, live chat, direct mail, marketing materials and social media. CRM systems can also give customer-facing staff detailed information on customers' personal information, purchase history, buying preferences and concerns. (adapted from https://searchcrm.techtarget.com/definition/CRM)

CRM data can get very extensive. Each metric you collect could carry some interesting information about your customers. But handling a dataset with too many variables is difficult. Learn how to reduce the number of variables in your data using principal component analysis. Not only does this help to get a better understanding of your data. PCA also enables you to condense information to single indices and to solve multicollinearity problems in a regression analysis with many intercorrelated variables.(adapated from Datacamp course: Marketing Analytics in R: Statistical Modeling)

EDA

Correlation matrix between Variables



We observe Intercorrelated variables in the bottom right area of the plot.

Standardization of all the variables

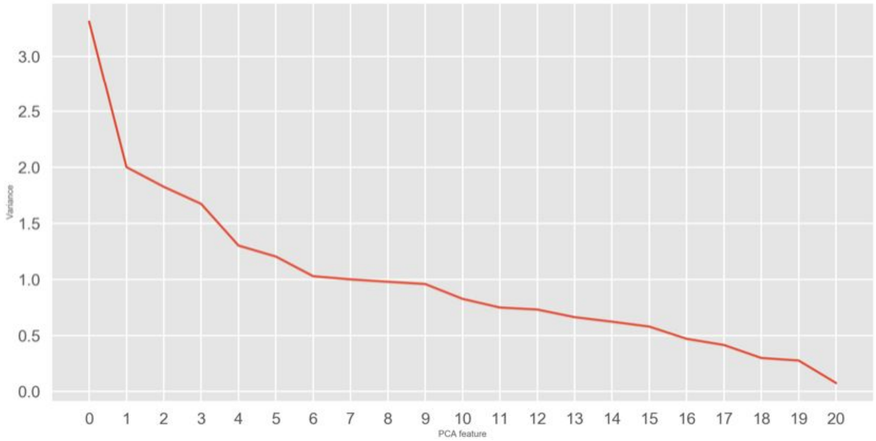
If some variables have a large variance and some small, PCA (maximizing variance) will load on the large variances. For example if you change one variable from km to cm (increasing its variance), it may go from having little impact to dominating the first principle component. If you want your PCA to be independent of such rescaling, standardizing the variables will do that. On the other hand, if the specific scale of your variables matters (in that you want your PCA to be in that scale), maybe you don't want to standardize (adapated from https://stats.stackexchange.com/questions/69157/why-do-we-need-to-normalize-data-before-principal-component-analysis-pca)

Our goal is dimension reduction. It's time to find out how many components you should extract. We can use several approaches to make this decision. Two methods that we will use here are

1. Scree plot
2. Kaiser-Guttmann rule

Scree plot

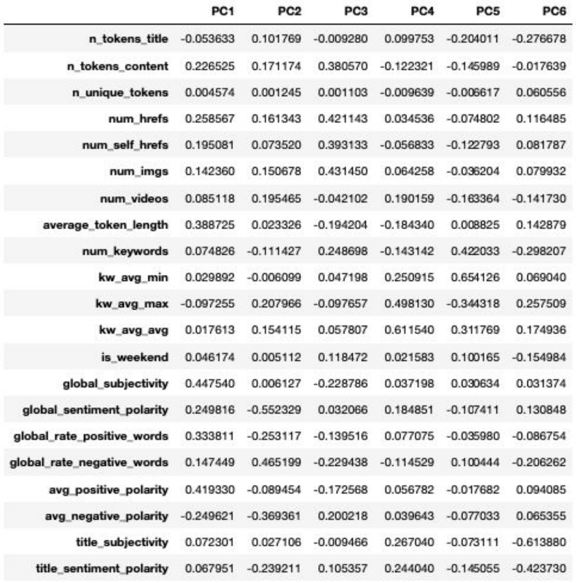
A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always displays a downward curve. The point where the slope of the curve is clearly leveling off (the “elbow) indicates the number of factors that should be generated by the analysis.



- 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)

The Kaiser rule is to drop all components with eigenvalues under 1.0 – this being the eigenvalue equal to the information accounted for by an average single item. The Kaiser criterion is the default in SPSS and most statistical software but is not recommended when used as the sole cut-off criterion for estimating the number of factors as it tends to over-extract factors.[20] A variation of this method has been created where a researcher calculates confidence intervals for each eigenvalue and retains only factors which have the entire confidence interval greater than 1.0 (adapted from: https://en.wikipedia.org/wiki/Factor_analysis#Older_methods)

PCA

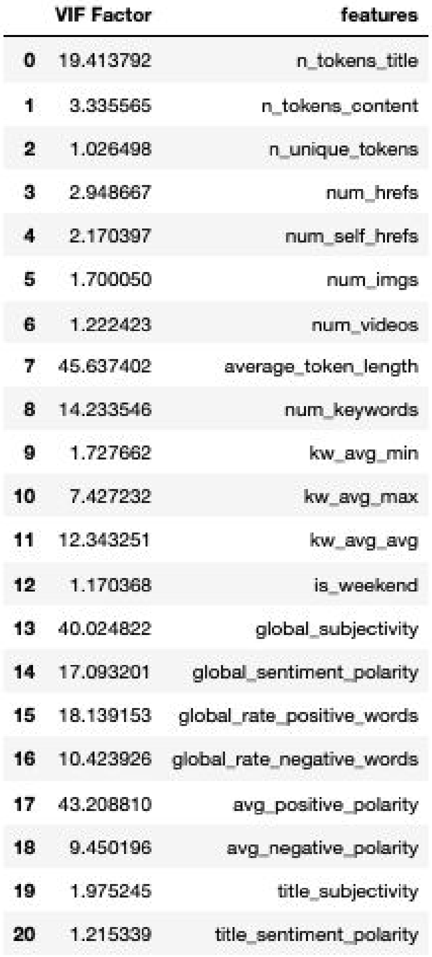


	PC1	PC2	PC3	PC4	PC5	PC6
n_tokens_title	-0.053633	0.101769	-0.009280	0.099753	-0.204011	-0.276678
n_tokens_content	0.226525	0.171174	0.380570	-0.122321	-0.145989	-0.017639
n_unique_tokens	0.004574	0.001245	0.001103	-0.009639	-0.006617	0.060556
num_hrefs	0.258567	0.161343	0.421143	0.034536	-0.074802	0.116485
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 Multiple Regression Analysis:

(A) Many Variables



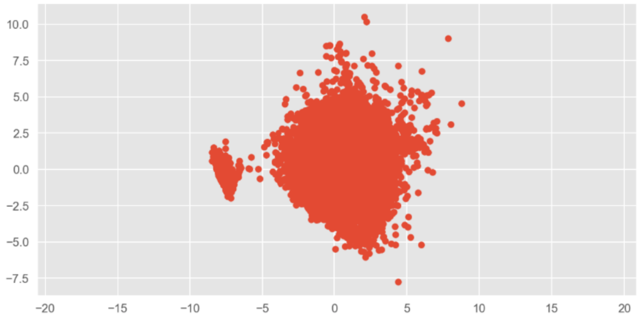
	VIF Factor	features
0	19.413792	n_tokens_title
1	3.335565	n_tokens_content
2	1.026498	n_unique_tokens
3	2.948667	num_hrefs
4	2.170397	num_self_hrefs
5	1.700050	num_imgs
6	1.222423	num_videos
7	45.637402	average_token_length
8	14.233546	num_keywords
9	1.727662	kw_avg_min
10	7.427232	kw_avg_max
11	12.343251	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

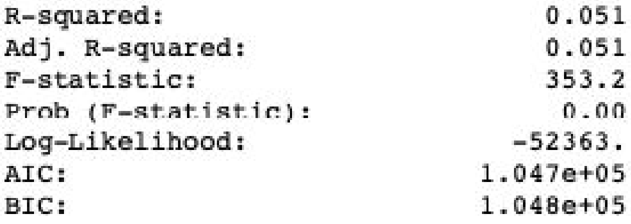
R^2 score = 0.08

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

(B) PCA components

Pearson 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
```

There is no correlation between PCA1 & PCA2

Multiple 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.

Conclusion of Part IV: 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 features/ 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.

(V) Project Summary of Market Analytics and Modeling Studies

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 which can be used are:

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