

Project Ideas: Capstone Project-2

Objective: Capstone Project 2	To implement machine learning tools in solving real life problems.
Project Idea (I)	<p>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.</p> <p>Data Sets: <u>Churn Data, Sales Data, Sales Data-months 2-4, survival data, Default data, News data.</u></p> <p>Questions:</p> <ul style="list-style-type: none"> • Which customers are valuable to your business? • Which customers will leave your business? • Survival Analysis. • CRM data Analysis.
Project Idea (II)	<p>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</p> <p>Data Set: a) HR data (2940, 4) b) Accident data (302,3), c) Survey data, d) Performance Data.</p> <p>Questions:</p> <ul style="list-style-type: none"> • 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)	<p>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.</p> <p>Data Sets: a) <u>Data Set from Kaggle</u> b) <u>From Datacamp</u> c) <u>Another one</u></p> <p>Questions:</p> <ul style="list-style-type: none"> • 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	<u>Churn Data, Sales Data, Sales Data-months 2-4, survival data, Default data, News data.</u>
Questions to be addressed	<ol style="list-style-type: none">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	<ol style="list-style-type: none">1) Linear Regression analysis2) Logistic Regression analysis3) PCA component analysis
Significance	<p><i>"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."</i></p> <p><i>KPI metrics: Sales Revenue, CPC, Online Marketing (ROI), Form Conversion Rates, Social media Search,</i></p>

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 four subsections: As a part of milestone report we will cover the parts which are highlighted below.

Market Analysis is divided into four subsections: As a part of milestone report we will cover the parts which are highlighted below.

- **(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 lifeline 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

Summary of Market Analytics and Statistical Modeling for Capstone Project-2

(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.

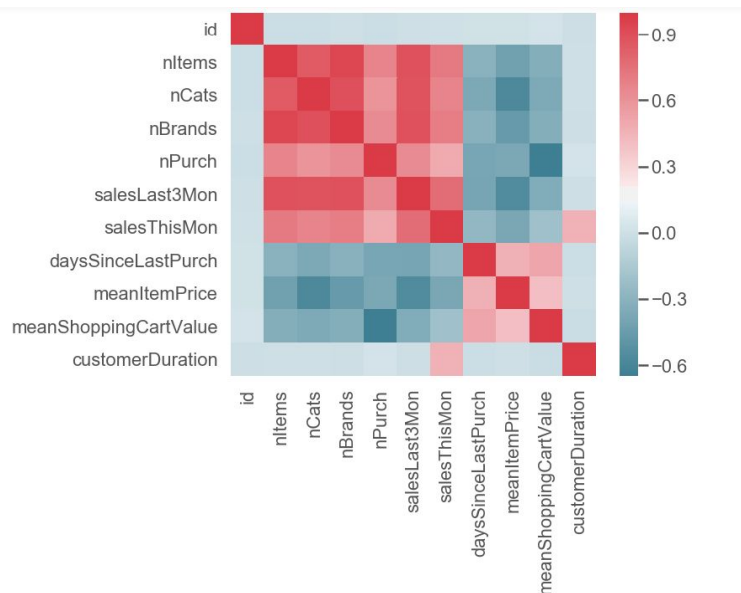
SalesData info:

```
<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

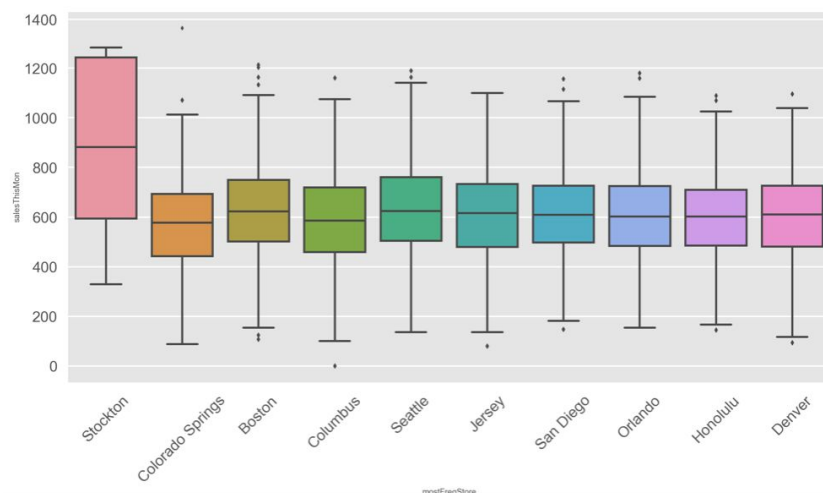
Correlation between the variables

We observe a correlation between 'salesLast3Mon' and 'salesThismon'

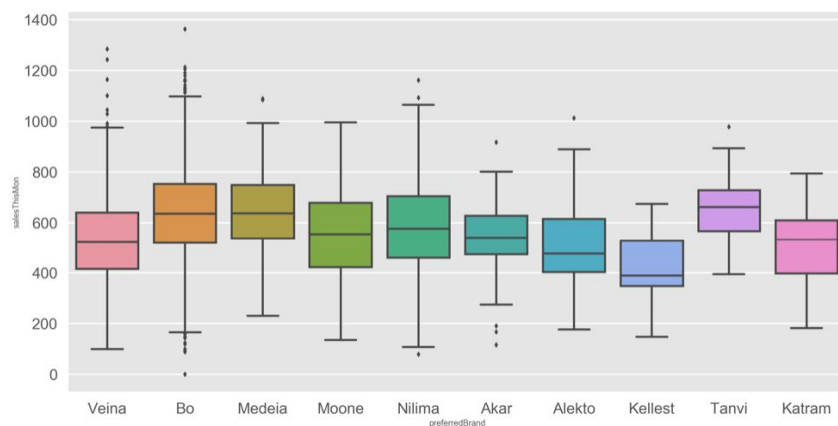


Further exploration and observing the correlation of sales over 3 months and categorical variables.

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

Regression Results

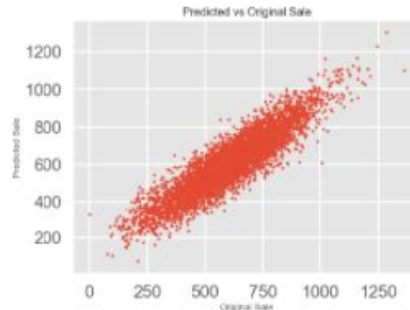
```
=====
R-squared:          0.825
Adj. R-squared:     0.824
F-statistic:        665.6
Prob (F-statistic): 0.00
Log-Likelihood:     -29521.
AIC:                5.912e+04
BIC:                5.936e+04
```

Multiple regression after dropping two variables:

mean vif = 5

Regression 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)

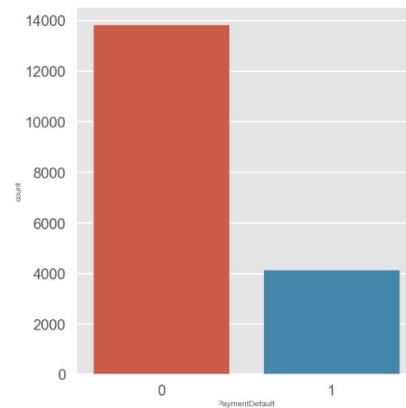
- Improved the regression score from 0.59 to 0.82 by including more variables.
- Multicollinearity problem when using all the variables, vif > 10.
- We took care of the multicollinearity by:
 - vif analysis and removal of variables with vif > 10
 - “preferredBrand” & “nBrands” had vif > 10 so we removed 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

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

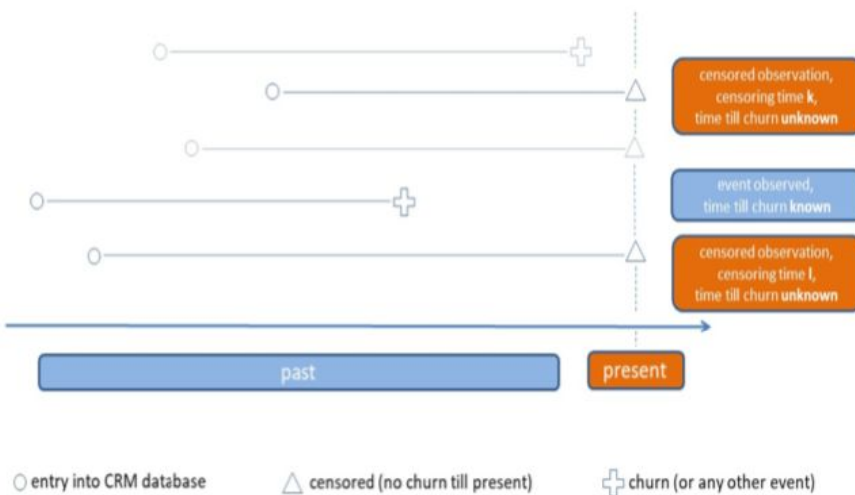
We will perform logistic regression model with all variables to model and predict the probabilities of customers who will default. This will be a part of final report.

(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 censored data or the missing data.

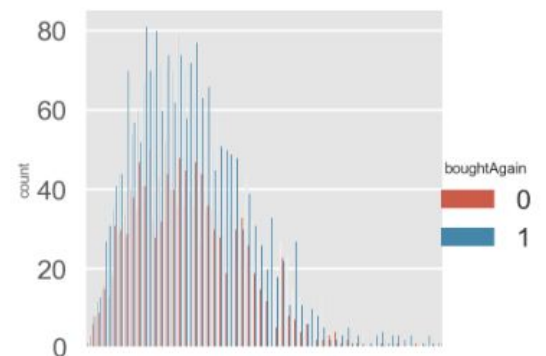
Advantages of survival analysis:

- 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.



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 observation: "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.

We will perform:

- 1) Survival analysis using Kaplan-Meier Fitting - without and with a categorical covariate.
- 2) CoxPH Model with Constant Covariates

(IV) CRM data analysis through PCA

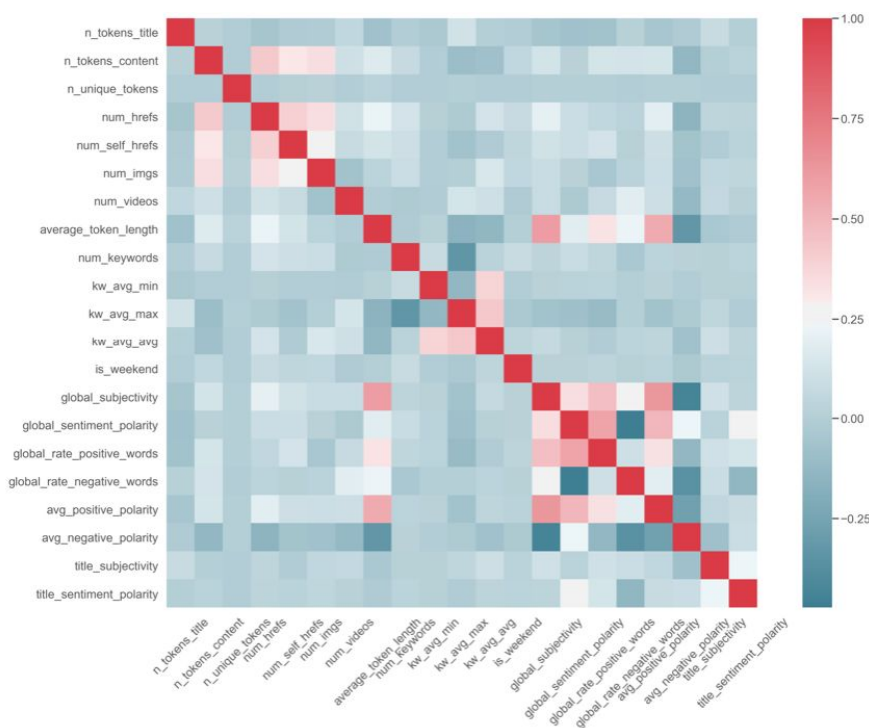
What IS CRM?

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. (adapted 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

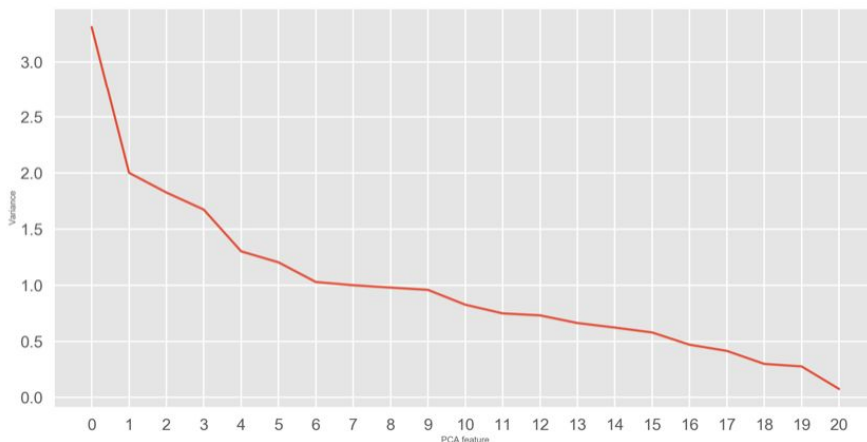
If some variables have a large variance and some small, PCA (maximizing variance) will load on the large variances. For example if you change one variable from km to cm (increasing its variance), it may go from having little impact to dominating the first principle component. If you want your PCA to 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 (adapted 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-Guttman 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-Guttman rule to drop variance below 1, suggests 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 corplot, but they split into two components.

We will perform multiple regression analysis with all the variable and then compare it to the analysis with PCA components.