IPCA
Data Mining

Prediction of Total Interactions of posts from a cosmetics brand on Facebook

SUMMARY

Keywords: CRISP-DM; Data Mining; Supervised Learning; Social media; Total Interactions; Linear Regression.

Summary

- Dataset is composed of 500 posts published by a well-known cosmetics brand on Facebook during 2014;
- CRISP-DM methodology applied;
- Purpose: to develop a total interactions prediction model that each post will generate;
- Total Interactions is the sum of likes, comments and shares of posts.

Introdução

Introduction

- The use of social networks is one of the most popular online activities in the world.
- The number of people on social media is expected to increase to almost 4.41 billion people by 2025;
- It is essential for brands on social networks (such as Facebook) to be able to improve their interaction strategies for their users and customize them to their target audience;
- Useful for Management: forecast the total interaction of users for each post published, from variables such as date / time, views and clicks on posts, among others.

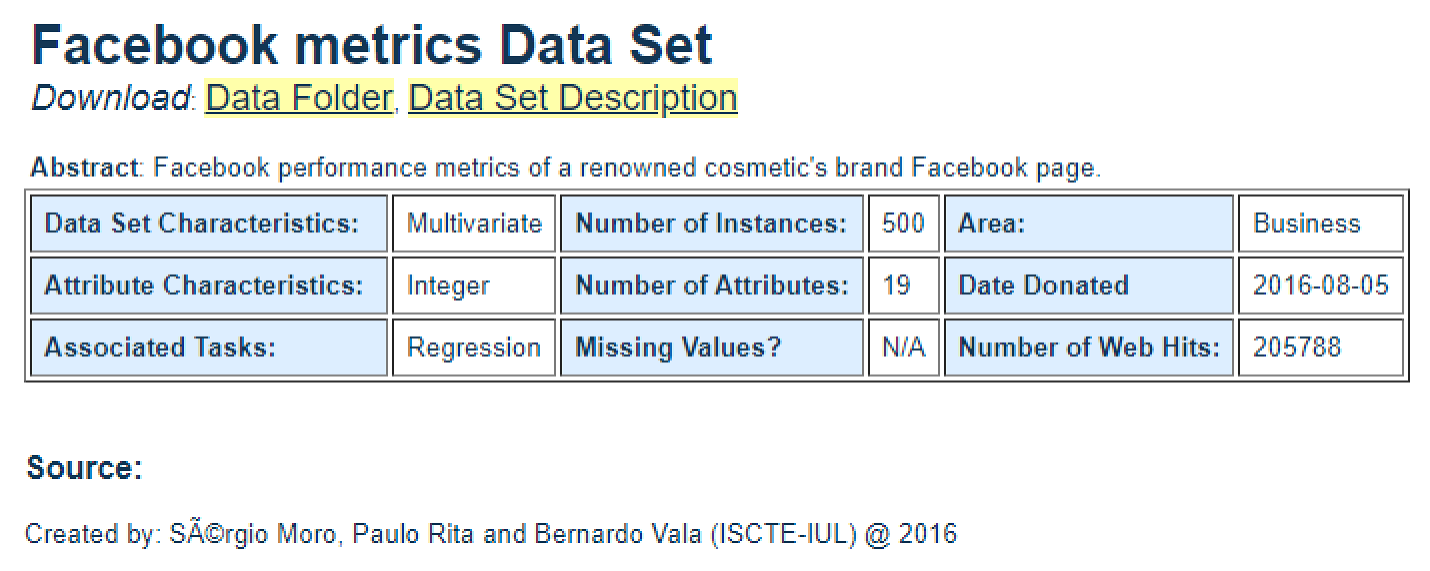
Materials and Methods

Utilized Data

Materials and Methods

Utilized Data

- The dataset was taken from the data file called "Facebook metrics Data Set", which is available at the UCI machine learning repository;



Facebook metrics Data Set

Download: Data Folder, Data Set Description

Abstract: Facebook performance metrics of a renowned cosmetic's brand Facebook page.

Data Set Characteristics:	Multivariate	**Number of Instances:**	500	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	19	**Date Donated**	2016-08-05
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	205788

Source:

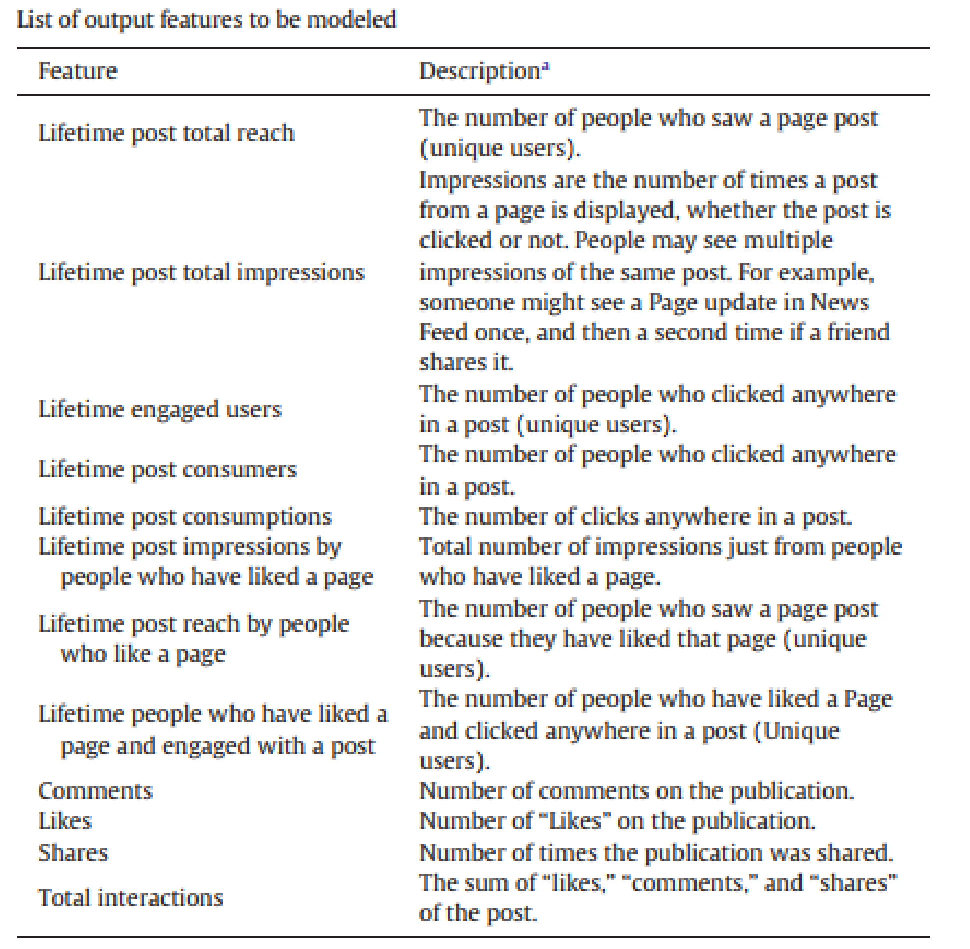
Created by: SÃ©rgio Moro, Paulo Rita and Bernardo Vala (ISCTE-IUL) @ 2016

- It includes 7 known attributes prior to post publication and 12 attributes that contribute to assess the posts post impact.

Materials and Methods

Utilized Data

- Description of the 12 attributes (Moro et al., 2016):



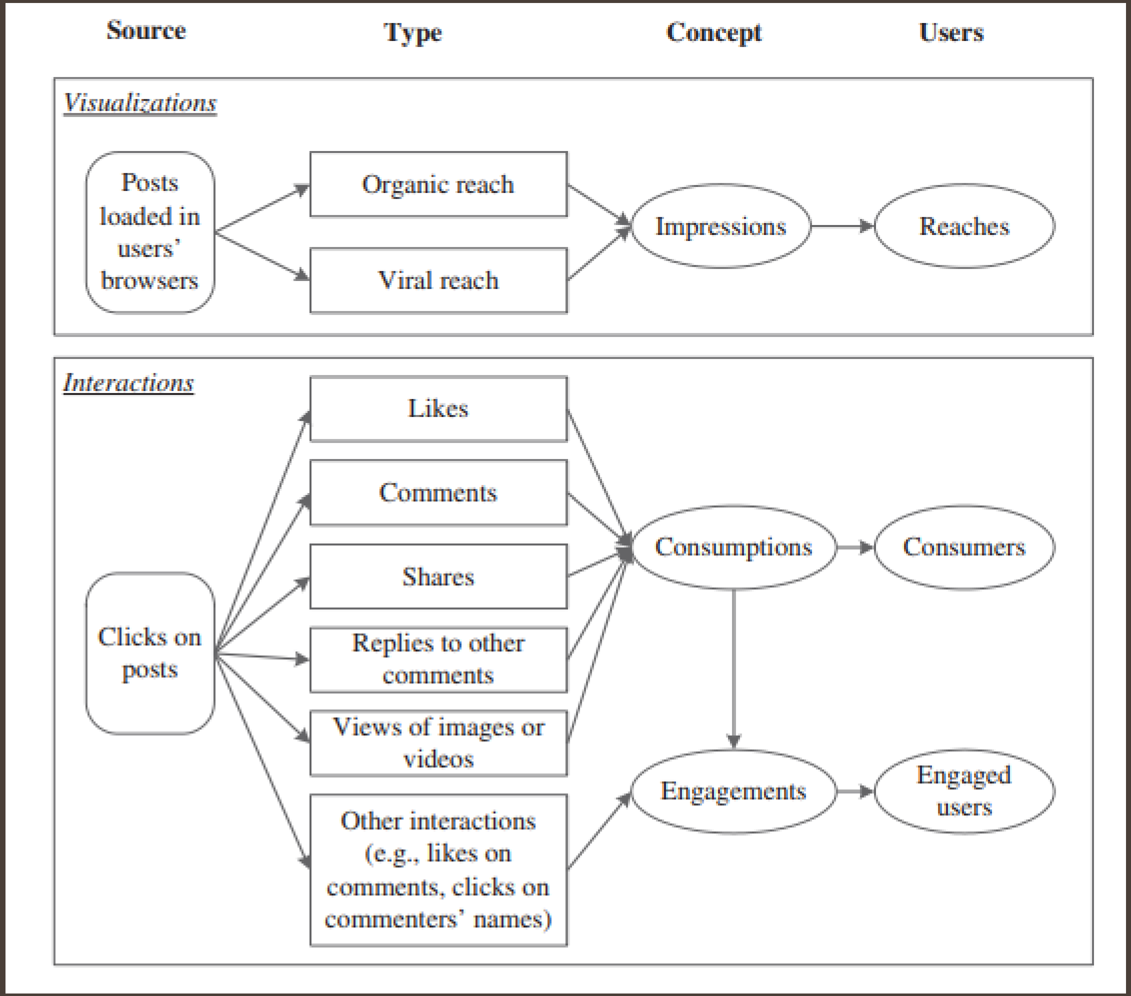
List of output features to be modeled

Feature	Description[a]
Lifetime post total reach	The number of people who saw a page post (unique users).
Lifetime post total impressions	Impressions are the number of times a post from a page is displayed, whether the post is clicked or not. People may see multiple impressions of the same post. For example, someone might see a Page update in News Feed once, and then a second time if a friend shares it.
Lifetime engaged users	The number of people who clicked anywhere in a post (unique users).
Lifetime post consumers	The number of people who clicked anywhere in a post.
Lifetime post consumptions	The number of clicks anywhere in a post.
Lifetime post impressions by people who have liked a page	Total number of impressions just from people who have liked a page.
Lifetime post reach by people who like a page	The number of people who saw a page post because they have liked that page (unique users).
Lifetime people who have liked a page and engaged with a post	The number of people who have liked a Page and clicked anywhere in a post (Unique users).
Comments	Number of comments on the publication.
Likes	Number of "Likes" on the publication.
Shares	Number of times the publication was shared.
Total interactions	The sum of "likes," "comments," and "shares" of the post.

Materials and Methods

Utilized Data

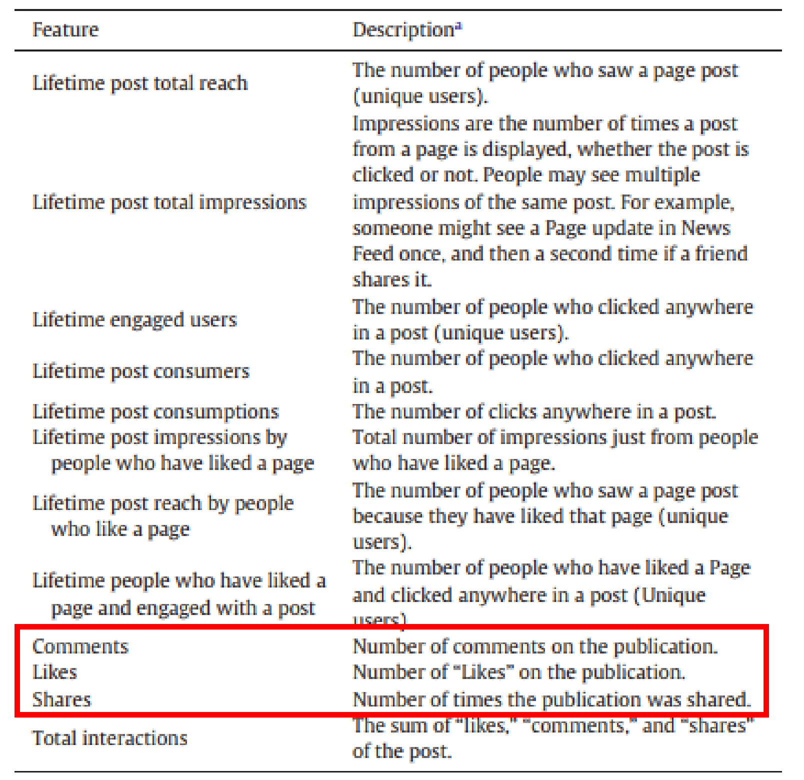
- These attributes relate to performance metrics of posts on Facebook (Moro et al., 2016):



Materials and Methods

Utilized Data

It would not be relevant to replicate this study, and the algorithm would learn to predict a variable based on the sum of the other variables:



Feature	Description[a]
Lifetime post total reach	The number of people who saw a page post (unique users).
Lifetime post total impressions	Impressions are the number of times a post from a page is displayed, whether the post is clicked or not. People may see multiple impressions of the same post. For example, someone might see a Page update in News Feed once, and then a second time if a friend shares it.
Lifetime engaged users	The number of people who clicked anywhere in a post (unique users).
Lifetime post consumers	The number of people who clicked anywhere in a post.
Lifetime post consumptions	The number of clicks anywhere in a post.
Lifetime post impressions by people who have liked a page	Total number of impressions just from people who have liked a page.
Lifetime post reach by people who like a page	The number of people who saw a page post because they have liked that page (unique users).
Lifetime people who have liked a page and engaged with a post	The number of people who have liked a Page and clicked anywhere in a post (Unique users).
Comments	Number of comments on the publication.
Likes	Number of "Likes" on the publication.
Shares	Number of times the publication was shared.
Total interactions	The sum of "likes," "comments," and "shares" of the post.

- Therefore, for this project we have eliminated the attributes "comments", "likes" and "shares", the sum of which would constitute the "Total Interactions".

Materials and Methods

Data Understanding

Materials and Methods

Data Understanding

- The database consists of 16 attributes:
- 10 numeric;
- 6 nominal

Atribute	Description	Atribute Type	Values
Page total likes	Performance	Numerical	
Type	Categorization	Nominal	Link, Photo, Status, Video
Category	Categorização	Nominal	Action (1), Product (2), Inspiration (3)
Post Month	Data	Nominal	1 a 12
Post Weekday	Data	Nominal	1 a 7
Post Hour	Time	Nominal	1 a 23
Paid	Categorization	Nominal	No (0), Yes (1)
Lifetime Post Total Reach	Performance	Numerical	
Lifetime Post Total Impressions	Performance	Numerical	
Lifetime Engaged Users	Performance	Numerical	
Lifetime Post Consumers	Performance	Numerical	
Lifetime Post Consumptions	Performance	Numerical	
Lifetime Post Impressions by people who have liked your Page	Performance	Numerical	
Lifetime Post reach by people who like your Page	Performance	Numerical	
Lifetime People who have liked your Page and engaged with your post	Performance	Numerical	
Total Interactions	Performance	Numerical	

Materials and Methods

Data Understanding

Descriptive statistics

- Numerical variables - descriptive statistics:

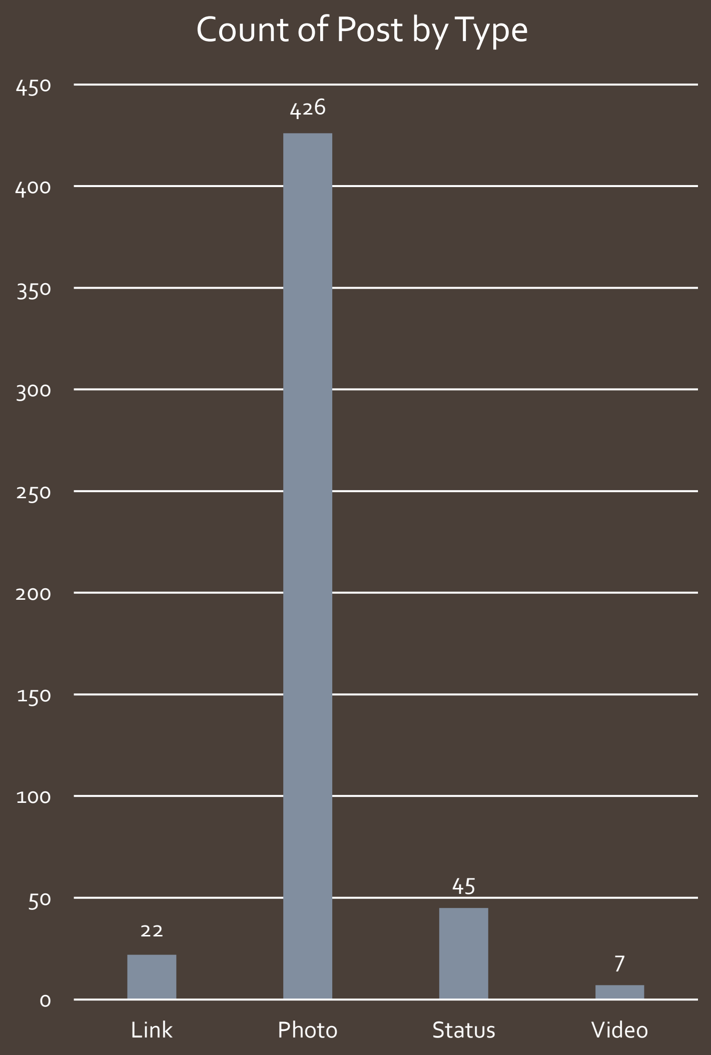
Atributes	Mean	Median	Mode	Standard Deviation	Minimum	Maximum	Count
Page total likes	123194,2	129600	136393	16272,8	81370	139441	500
Lifetime Post Total Reach	13903,4	5281	5280	22740,8	238	180480	500
Lifetime Post Total Impressions	29585,9	9051	4372	76803,2	570	1110282	500
Lifetime Engaged Users	920,3	625,5	537	985,0	9	11452	500
Lifetime Post Consumers	798,8	551,5	182	882,5	9	11328	500
Lifetime Post Consumptions	1415,1	851	431	2000,6	9	19779	500
Lifetime Post Impressions by people who have liked your Page	16766,4	6255,5	3675	59791,0	567	1107833	500
Lifetime Post reach by people who like your Page	6585,5	3417	1640	7682,0	236	51456	500
Lifetime People who have liked your Page and engaged with your post	610,0	412	403	612,7	9	4376	500
Total Interactions	212,1	123,5	75	380,2	0	6334	500

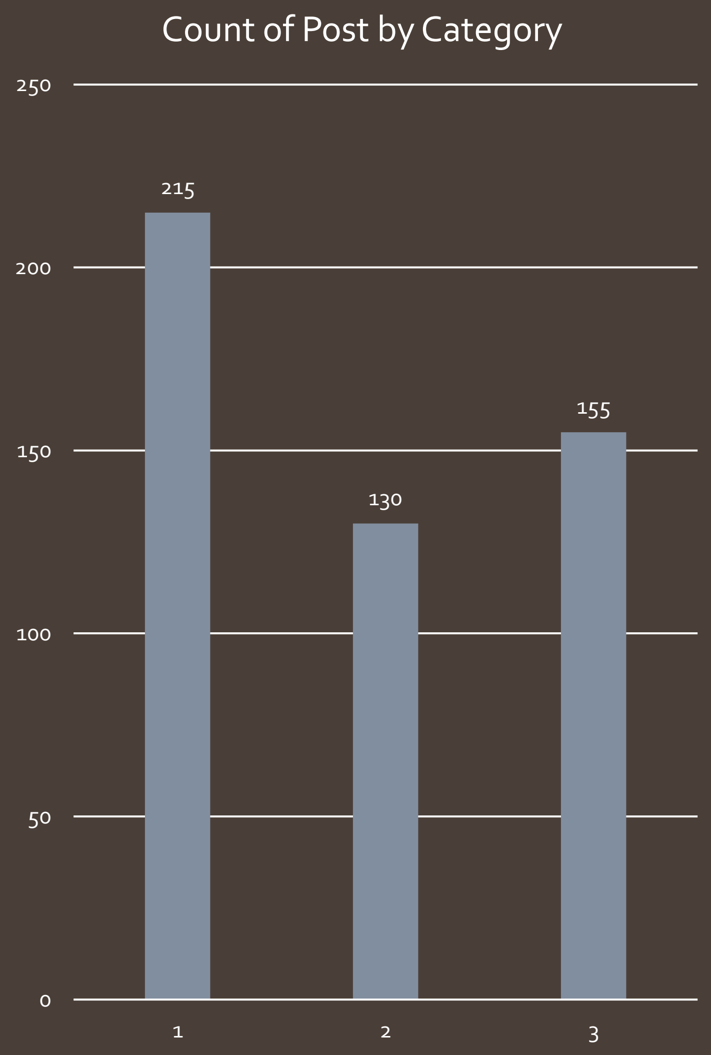
Materials and Methods

Data Understanding

Histograms

Nominal variables - histograms



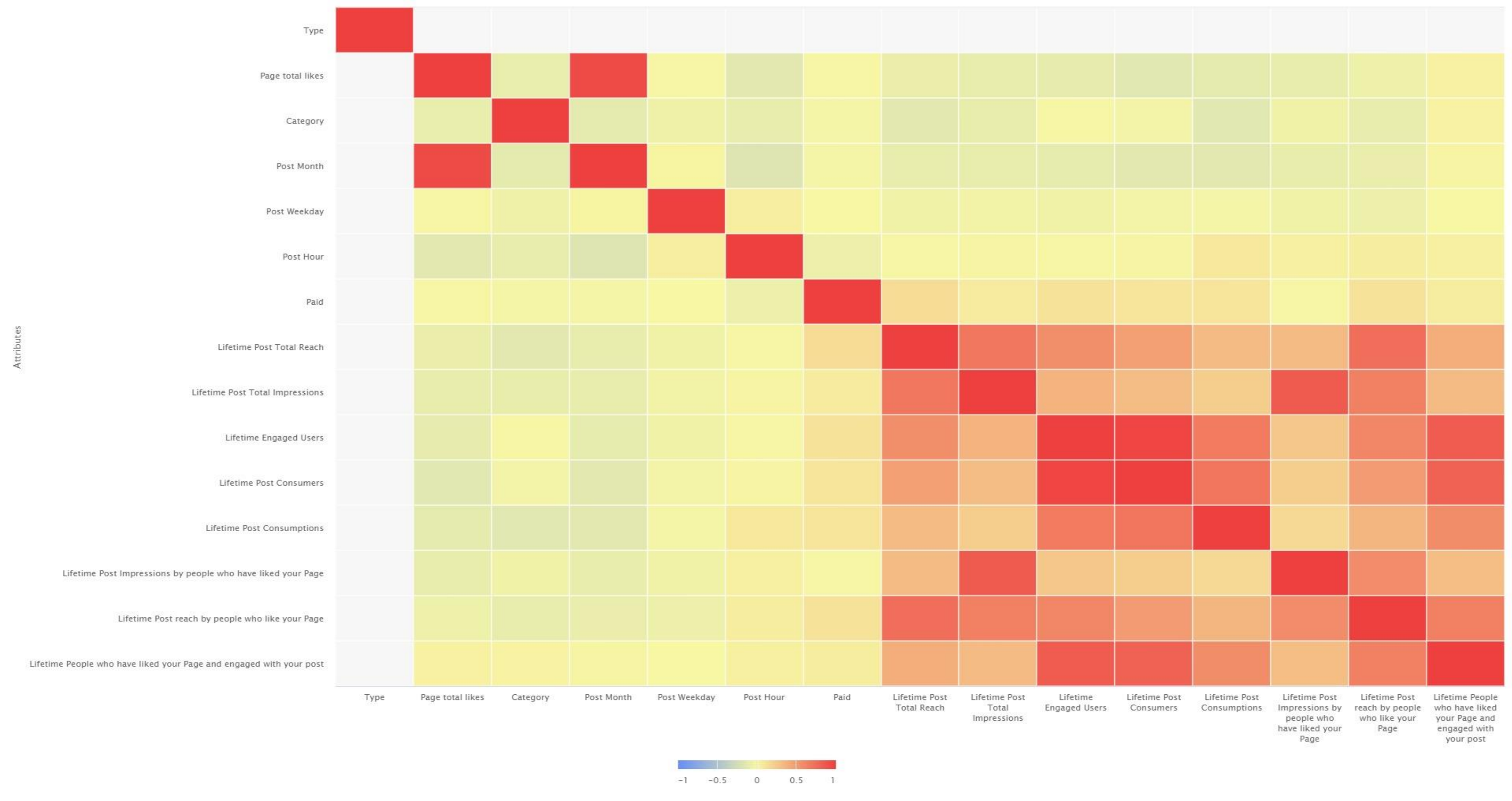


Materials and Methods

Data Understanding

Correlations

- The correlation matrix shows the correlations between all dataset variables:
-



Materials and Methods

Data Understanding

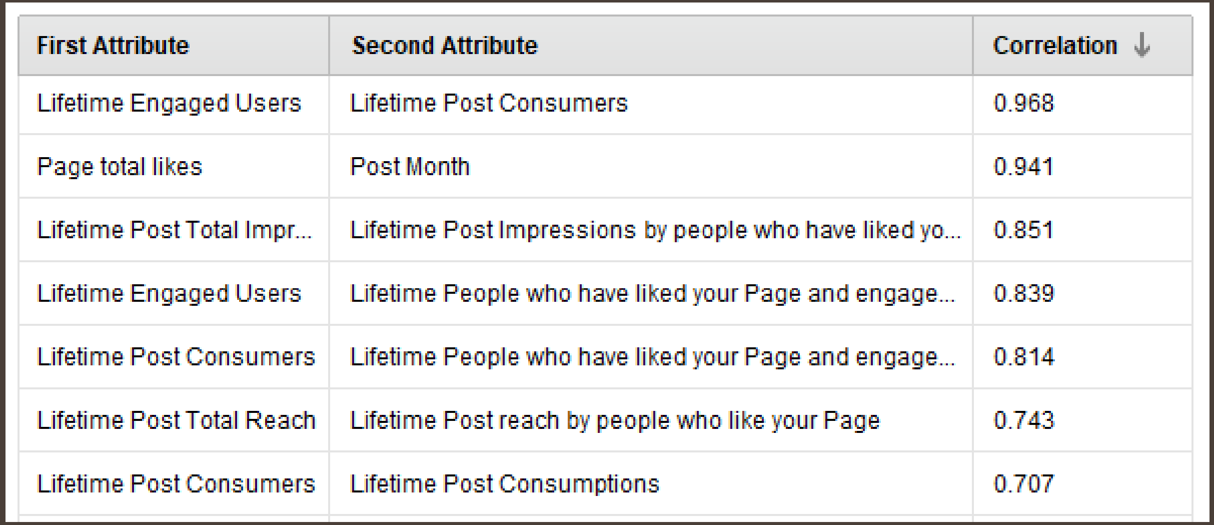
Correlations

- Strong positive correlations:

1) Lifetime Engaged Users vs Lifetime Post Consumers;

which would be expected, as they both relate to the number of clicks in a post (single users and total clicks, respectively).

2) Page total likes and the month of publication of the post.



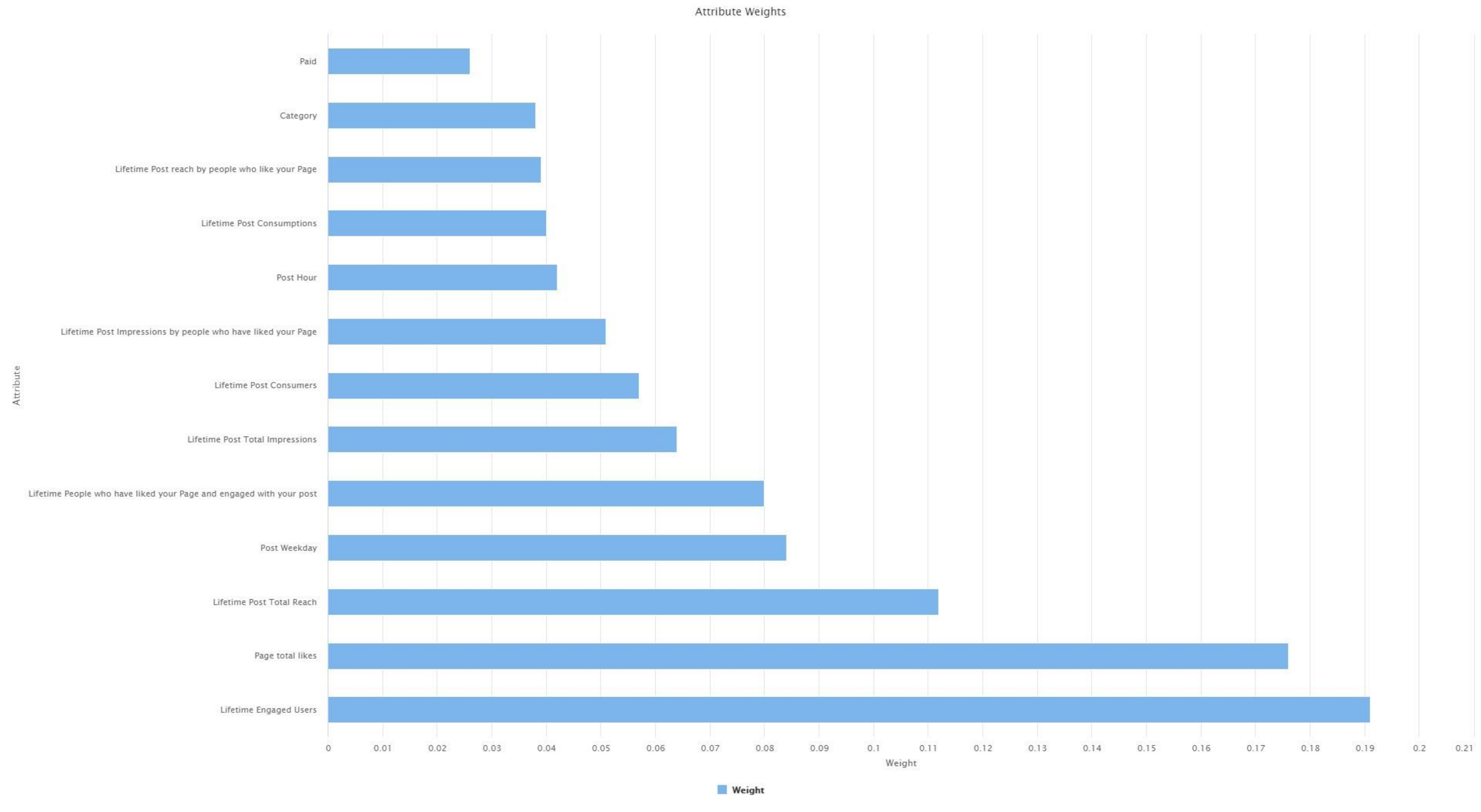
First Attribute	Second Attribute	Correlation ↓
Lifetime Engaged Users	Lifetime Post Consumers	0.968
Page total likes	Post Month	0.941
Lifetime Post Total Impr...	Lifetime Post Impressions by people who have liked yo...	0.851
Lifetime Engaged Users	Lifetime People who have liked your Page and engage...	0.839
Lifetime Post Consumers	Lifetime People who have liked your Page and engage...	0.814
Lifetime Post Total Reach	Lifetime Post reach by people who like your Page	0.743
Lifetime Post Consumers	Lifetime Post Consumptions	0.707

Materials and Methods

Data Understanding

Target Variable

- By applying the Decision Trees model in RapidMiner, it is visible that the variables which more influence Total Interactions are (in descending score order):



Materials and Methods

Data Understanding

Target Variable

- The most significant interactions in posts (comments, likes, shares) are associated with:
- people who clicked on the post;
- page total likes;
- people who saw the post of the page.
- These factors have more influence in Total Interactions than e.g. the Type and Category of the post.

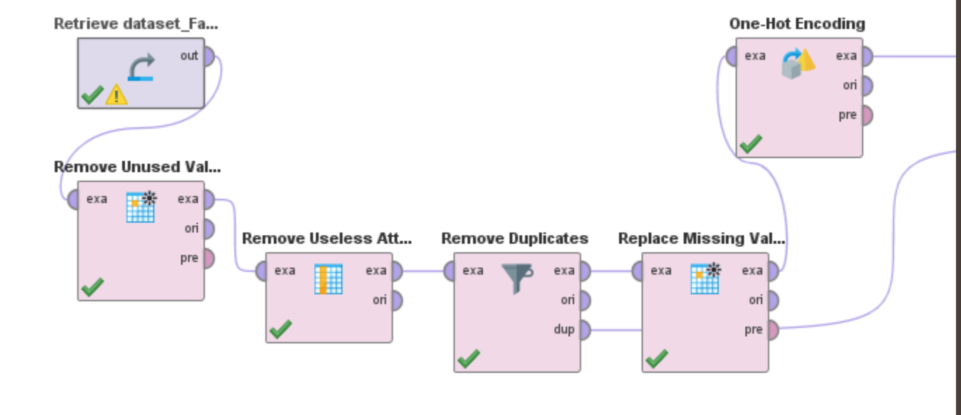
Materials and Methods

Data Preparation

Materiais e Métodos

Data Preparation

- The data cleaning involved the following operators:

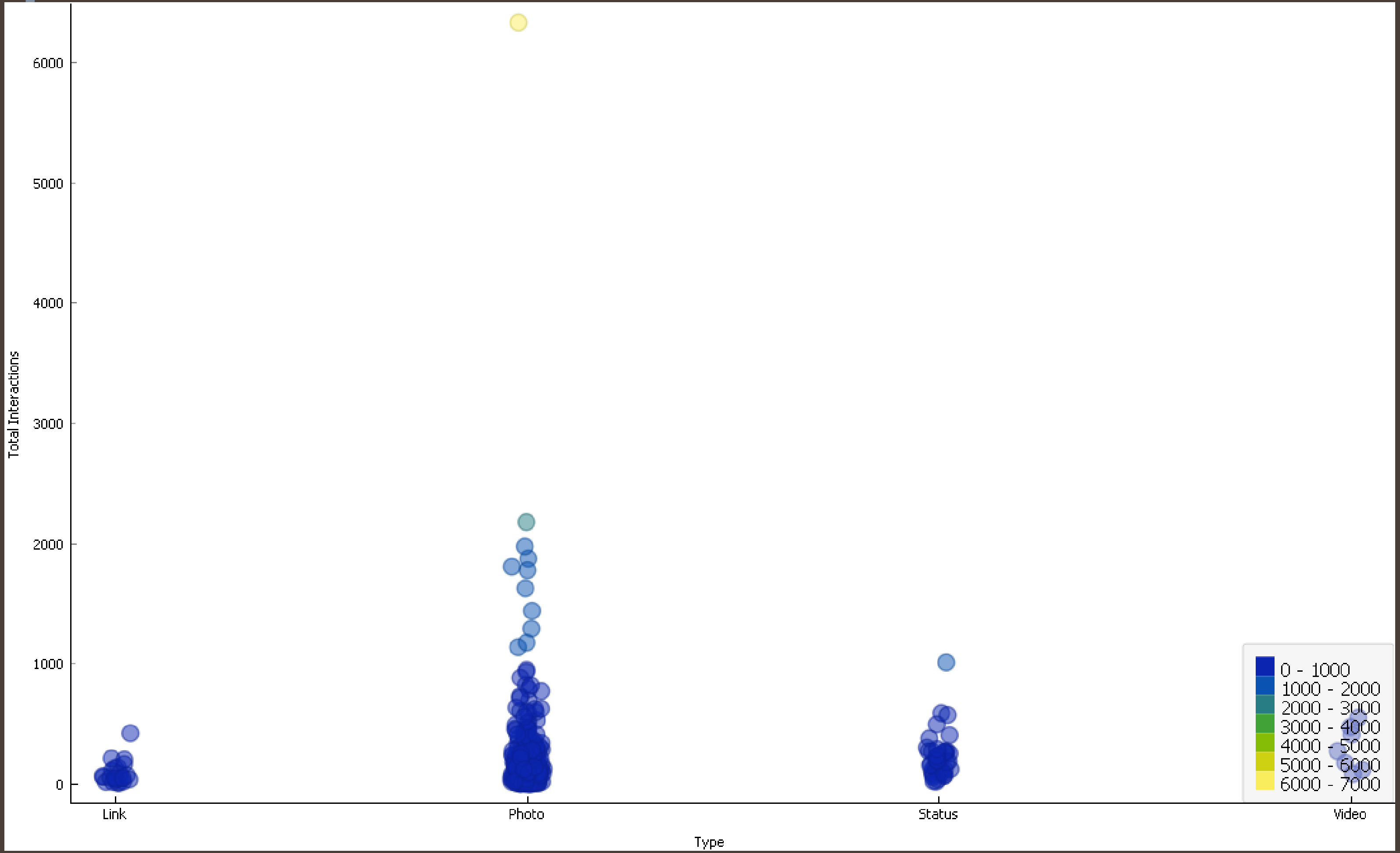


- - remove unused values;
- - remove useless attributes;
- - remove duplicates.
- Missing values: 1 missing element in the "Paid" column.
- Conversion of polynomial variables to numerical: variable "type“; the one-hot encoding technique was used (transitioning from discrete to numerical variables in the "type" attribute).
- Throughout the Data Reading, several outliers were identified, which will be filtered on a later stage.

Materials and Methods

Data Preparation

- Outlier in the scatter plot of total interactions for the post type:



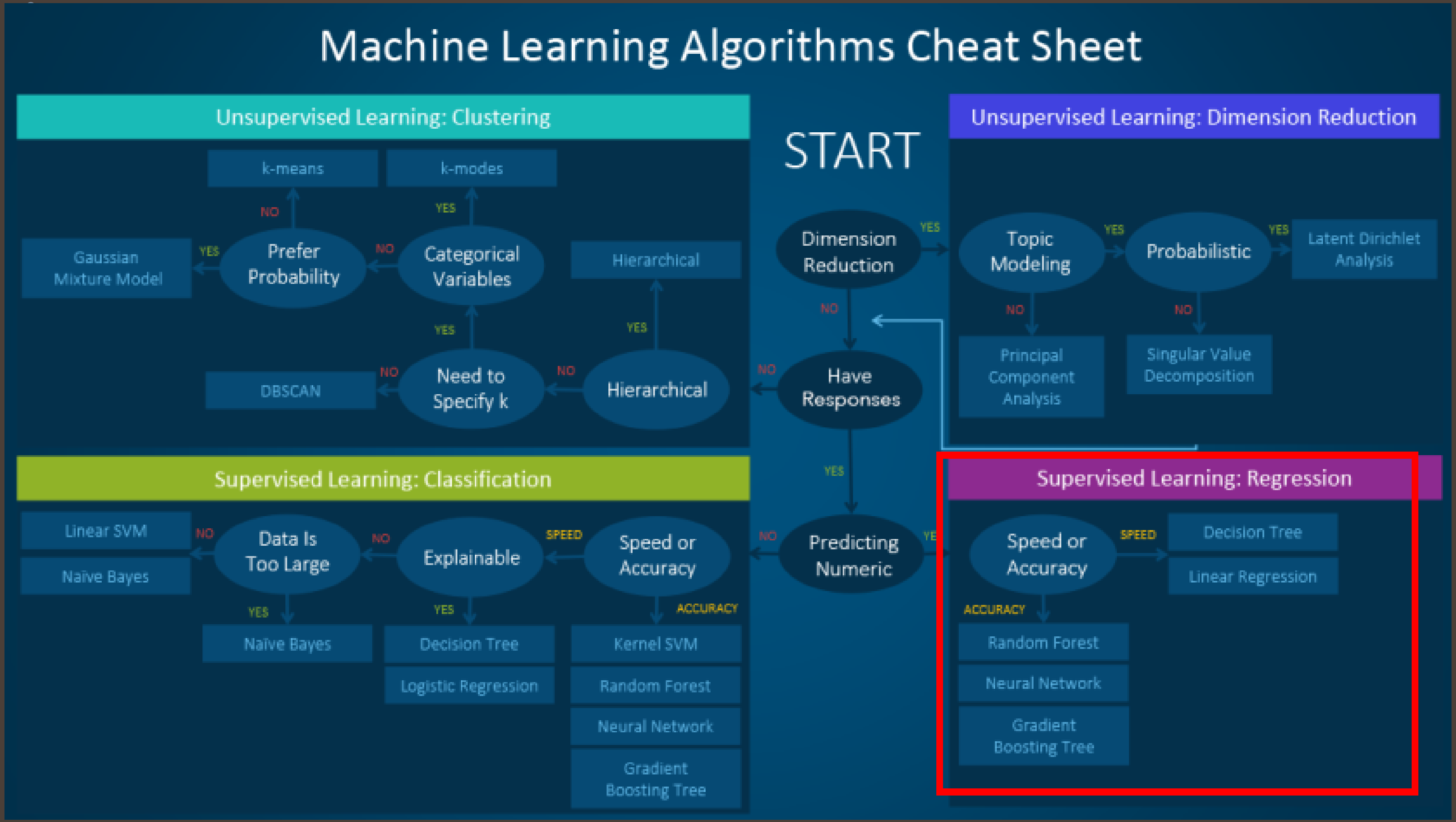
Materials and Methods

Modeling

Materials and Methods

Modeling

- 3 algorithms were selected based on the criteria of the guide published by SAS and also on previous works.



Materials and Methods

Modeling

Since the objective is to predict a numerical variable (through a regression analysis) and with supervised learning, the following algorithms were applied:

Linear Regression;

Decision Tree;

Random Forest;

Vote (LR + DT).

Materials and Methods

Evaluation

Materials and Methods

Evaluation

Evaluation metrics:

As in regression, we're predicting continuous values (decimal numbers), the evaluation metrics will focus on measuring the error between forecasted and real values.

The 2 regression metrics that were considered are:

- Mean Absolution Error (MAE)
- Root Mean Squared Error (RMSE)

Materials and Methods

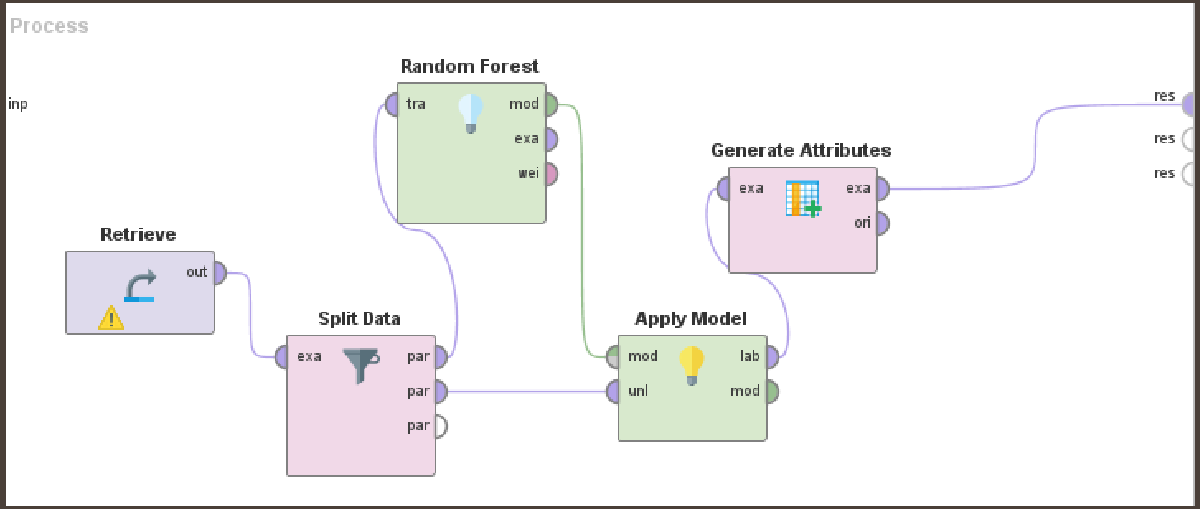
Evaluation

- Sampling techniques:
- - Split: the column is divided into a fixed percentage (p) for training and another (1-p) for testing.
- It was decided in this case to use the ratio 0.75 / 0.25.
- - Cross Validation: in 'traditional' cross-validation the total examples (P) are divided into mutually exclusive k subsets (P1, P2,...,Pk) with approximately equal sizes (k-folds).
- In the current example, it was applied a set of 10 folds.

Results / Discussion

Results / Discussion

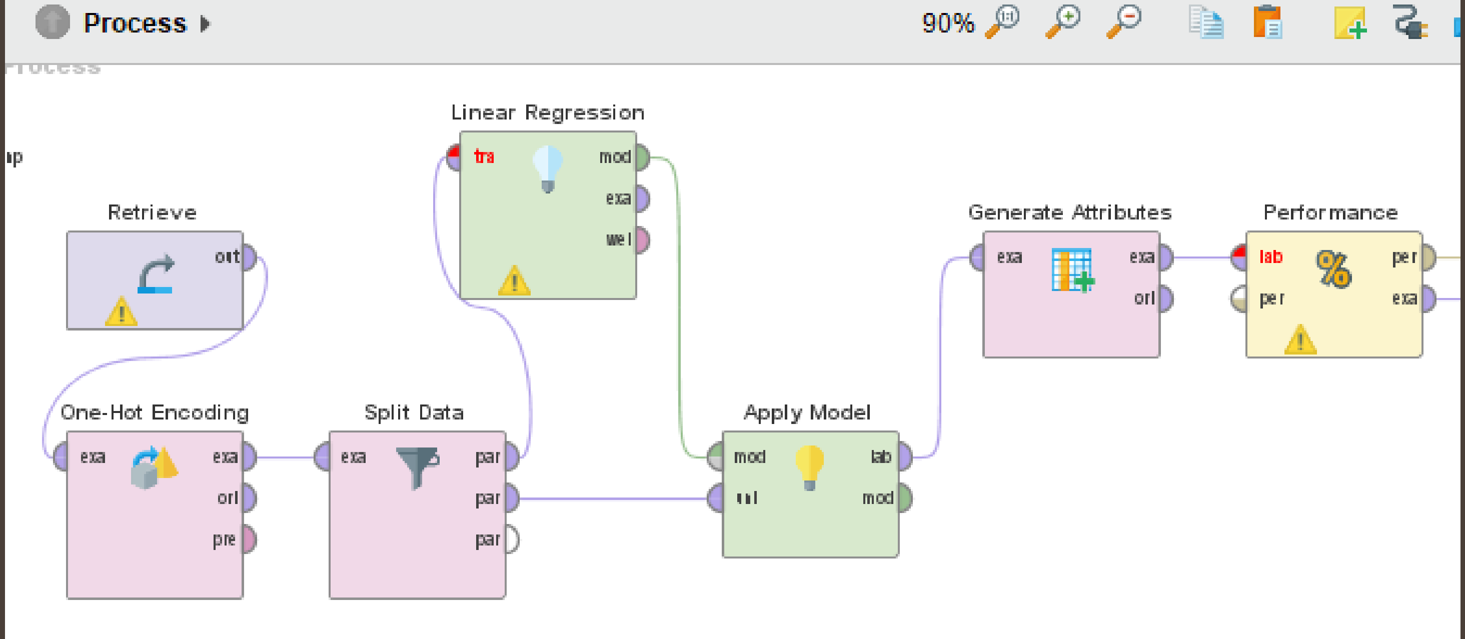
- With the applied models we were able to generate a new column called "prediction (Total Interactions)" to predict the target variable “Total Interactions”:



		Column Difference	
Model	Min	Max	Average
Linear Regression	0	273	35,6
Random Forest	1	896	77,2
Decision Tree	0	1426	82,1
Vote (LR + RF)	0	280	43,9

Results / Discussion

- Performance measurement to determine MAE:



Results / Discussion

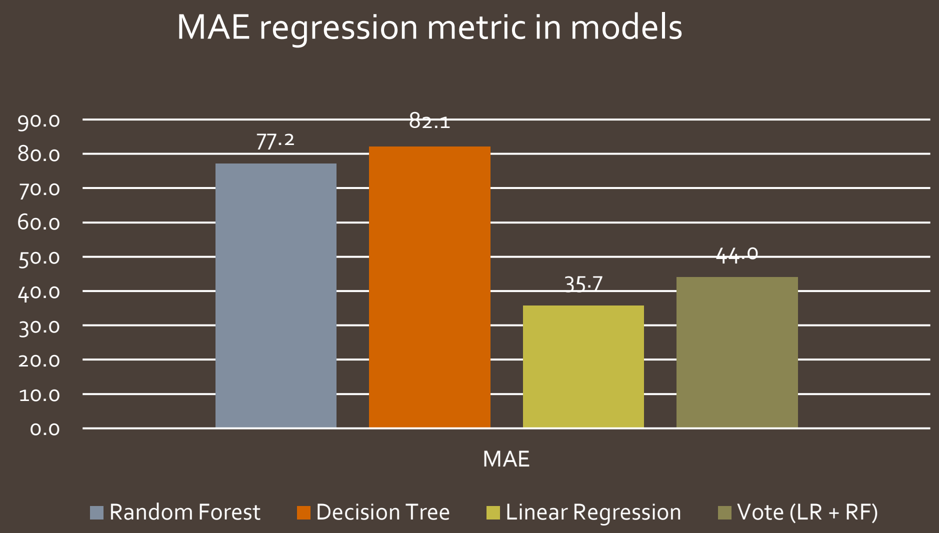
Sampling techniques obtained with the Random Forest algorithm:

	Regression metrics	
Sampling techniques	MAE	RMSE
Split	**77,2**	**147,1**
Cross Validation	96,2	236,7

- There is a significant difference between The MAE and the RMSE, which means that in the dataset there will be extreme errors.
- RMSE is a metric that penalizes large errors (that is, lines where the algorithm has significantly failed). Thus, it is greatly affected by the existence of outliers.

Results / Discussion

- MAE results for the 4 applied models :



- The Linear Regression algorithm turned out to be the model with better performance, that is, with a lower MAE.

Results / Discussion

The following techniques were applied to reduce the value of the MAE:

- normalization (3 variables were selected: post hour, post weekday and post month), since they concern the date/time category and its interpretation is possible even after normalization;

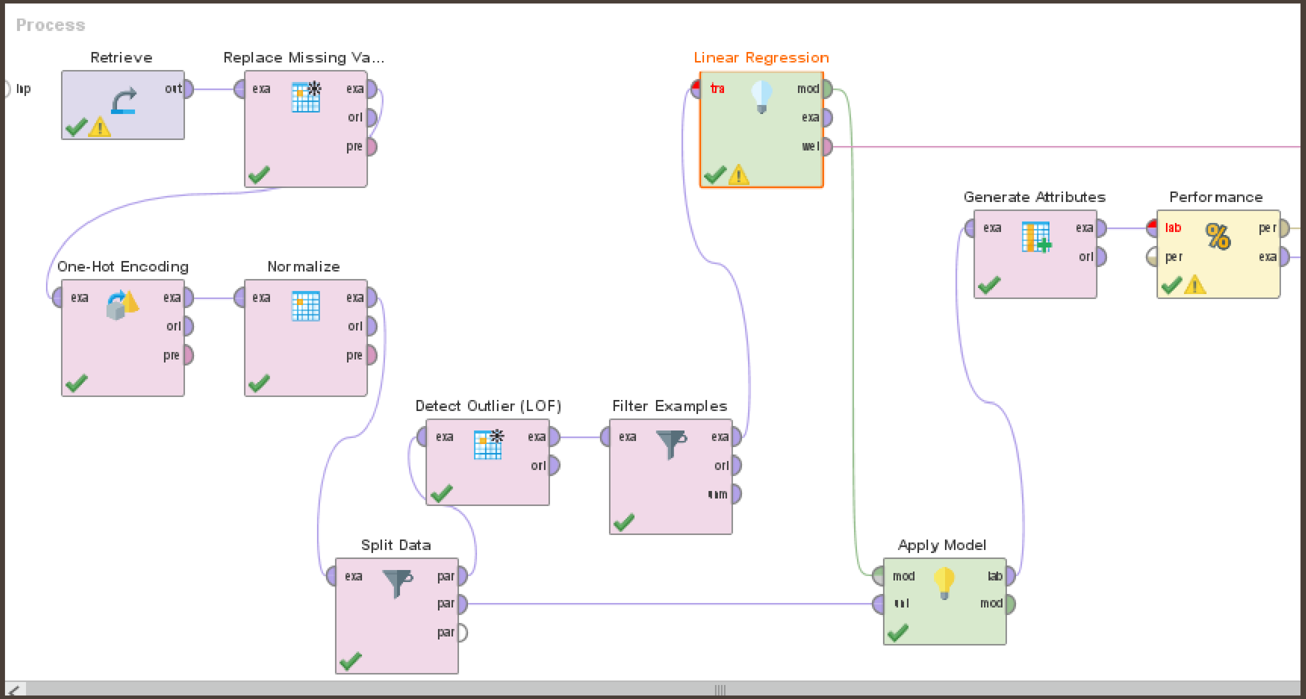
-detection and filtering of outliers (< 3, < 2 and < 1).

Operator	MAE
Normalize (hour, day, month)	35,5
Remove Outliers (> 3)	38,1
Remove Outliers (> 2)	39,8
Remove Outliers (> 1)	**18,8**

- With filter outliers (> 1), the MAE drops to almost half of its value.
- However, there is a drastic reduction in the number of elements (from 500 to 24 records).

Results / Discussion

In order to validate the outliers filter > 1: outliers were eliminated in the training data (75%), maintaining the test data set (25%) with outliers, and by applying split validation:



If there is no major change in the MAE, then the best filter is < 1

Results / Discussion

- Validation results of outlier filters:

Test plot with outliers	MAE	Variação
Filter < 1 outliers test plot	151	132,20
Filter < 2 outliers test plot	47,6	7,80
Filter < 3 outliers test plot	**36,2**	**-1,90**

- As the MAE in the < 1 and < 2 filter has increased and in the case of the < 3 filter has reduced, it is concluded that this filter is the most reliable.

- From the Deployment point of view, outlier filtering should be rechecked, as a tighter filter can reduce the ability to predict new records with outliers from dataset without outliers.

Conclusion

Conclusion

- Linear Regression was the model that performed best and is also easily interpreted by humans.
- The large difference between the MAE and RMSE values described in item 4 indicates the existence of large errors in some lines and small in other lines;
- The models' results were thus greatly affected by the existence of the outliers;
- Filtering these outliers to values below 1 substantially dropped the MAE value;
- However, scouting the robustness of outlier filters is also important, as outliers are part of all datasets and should be predicted (even if this implies increasing the value of the MAE a bit).

END

Thank you!