Analysis of Webshop Data

Fundamentals SCC.460 Group Project - Group 5

Chiotis, K.(35278597), Clayden, W. (35190586), Frey, M.-A.- (35313329),
Mishra, P.(35390350), Werawatganon, E.(35337880), Zhang, Z.(35303577)

13. January 2020

The data used in this project was kindly donated by the N Brown Group plc data science department.

For creating this report, the work was split between members as shown below.

Analysis

- Pre-Processing – Kyriakos
- Exploratory data analysis – Everyone
- Visualization of web page flow – William, Kyriakos, Marc
- Classification Modeling – Pritam

Writing

- Introduction – Will
- Pre-Processing – Max
- Question 1 – Will, Kyriakos
- Question 2 – Kyriakos
- Question 3 – Pritam, Marc
- Clustering – Jack
- Conclusion – Marc

1 Introduction

NBrown are a massive online retailer operating in the UK and Ireland as a prominent Figure in many areas of fashion as well as home wares. Their focus is on inclusivity with wide ranges of clothing for all sizes and to be future focused emphasising caring for their customers and the planet. In their words "We serve our customers through a family of trusted fashion brands".[1] This keen interest to benefit customers and make a friendly online purchasing environment is reflected in all aspects of their stores, from products designed with inclusivity in mind, to their financial services they offer. These services enable "credit" to be offered, so that customers can spread the cost of products to make them more affordable. Similarly, NBrown also has a "try before you buy" mechanism. These services are provided directly by NBrown.

The analytic data they provided us with, is web-traffic data to be analysed such that inference and suggestions could be made that would be both beneficial to the user-base and similarly the company. With recent ruling regarding credit and rent to own schemes affecting the profitability of the financial services in the company, a growth of the user base, an increase in registered members as well as improvements in the conversion rate will help account for the losses from the legislative changes (conversion - when a session ends with a purchase).

Knowing this information we came up with some questions to answer:

1. How different factors relate to conversion rate and can any inferences or hypotheses be drawn from the findings?
2. How do people typically move around the websites?
3. Is it possible to create a model that can predict whether someone is likely to buy something?

The questions are explored and explained in the Methodology section. In the results section, each question is addressed and some inferences or recommendations are made. We think that these questions would be most beneficial to NBrown to identify: The first question allows for identification of possible areas of investment or further inference with more or different data. The second question allows for understanding of how people navigate websites and whether there are any patterns observed in this that are unnecessary. Finally the third question seeks to consider whether predicting a checkout is possible when someone arrives at the page and as they navigate through with the hope that knowing this live could be beneficial in some way.

2 Methodology

2.1 Research Strategy

The research strategy began individually looking at the data and determining whether there were any imperfections in the data set that was supplied by NBrown. The data consisted of over 92000 unique rows with 24 variables and upon importing the data some of us found errors that other group members imports didn't generate. A standard dataset was made that everyone could use without importing errors. Following from this it was determined necessary to pre-process the data and then divide into subgroups to address the research questions. The approaches taken for each vary as a mix of Exploratory Data Analysis and statistics, flow diagrams and both Machine Learning Models.

2.2 Pre-processing

The first stage of data analysis required pre-processing to prepare data for further steps. In this part, we will conduct various approaches of pre-processing, consisting of data cleansing, data extraction, data transforming and feature scaling.

For **data cleansing**, we have checked missing values, such as Null/Nan, and remove some row that contain these values. Moreover, the rows containing spam visiting and some duplicate session have been deleted. Besides, we changed string values, capital and small character, to be the same. In addition, some features that are uncorrelated will be removed, as well as the feature containing response variable (SESSION END STATUS) will be separated before conducting clustering. This dependent variable will be used as checking labels to measure the model performance.

As for **data extraction**, some important features have been extracted from the data such as total time of the session, total sequence of pages visited as well as the timestamp. All these features were generated for classification and clustering purpose.

Besides, categorical features have been converted to binary and numeric features, such as status that people will check out or not. Furthermore, for continuous features, we conducted standardisation to scale data equally for comparability. These techniques were the prerequisites for the next stage which were classification and clustering.

2.3 Methodology of Question 1

The first question we sought to investigate did initially receive far too much manpower relative to what was necessary to objectively complete the task and answer the question "How to different factors relate to conversion rate and can any inferences or hypotheses be drawn from the findings?". The initial approach was to collectively research specific areas of the data and find some exploratory analysis. Whilst successful this was not unnecessary as a few areas ultimately were not required. The data was manipulated in RStudio using the dplyr library to generate multiple plots. Each plot featuring a categorical variable from the table below (Figure 1) then

observing the checkout rate of each factor level of what was found.

Variable	Variable Description
WEB_SITE_NAME	The Website URL of each website visited. The conversion rate only was counted for the website associated with the purchase (i.e. Checkout Completed)
DATE_AND_TIME_OF_PAGE_VIEW	The date the user visited any one of the websites. This was split into year, day of the year and day of the week.
DEVICE_TYPE	The device type the session was on (Mobile, Tablet or PC)
DEPARTMENT_DESC	The department type e.g. womenswear or footwear. This specifically is any department that was visited that has a session end status of Checkout complete

Figure 1: Variables used and their descriptions

This showed an exploratory analytic view of the behaviours of the data. Then selecting the variables that were deemed of best inferential value box-plots then ANOVA with hypothesis tests were conducted to determine whether the observed trends were valid to comment on.

2.4 Methodology of Question 2

The second research question was formulated under the aspect of understanding the navigation process of website visitors. It could be divided in sub questions and focus on different variables to cover a range of case studies. The variables regard the navigation of PC users against the mobile users and what is the sequence of visited pages for completed checkouts against abandoned bags and orders. The answers on this research could provide valuable information about difficulties on user navigation experience and defective structural pages. The results were delivered using exploratory data analysis and particularly a chord diagram. A chord diagram creates links between the pages under a web session with respect on the ordinal sequence of visited pages. Some page types that had very few connections were omitted to prevent over complication of the graphic. Thus, the method focuses on the connections between the Home, Navigation, Product View, Search Results and View Bag pages.

2.5 Methodology of Question 3

2.5.1 Clustering

(1) K-Prototype clustering

To divide the data into checkout and no checkout groups, you should first analyze the data variables. This data contains both categorical and continuous variables. Therefore, k-means, K-mode and hierarchical clustering cannot be used to cluster the data, because K-means and hierarchical clustering are only applicable to data that are all continuous variables. The reason is that in the process of using these two clustering methods, the Euclidean distance between samples needs to be calculated, but for categorical variables, the Euclidean distance has no meaning. Similarly, for k mode, the Hamming distance needs to be calculated, and the Hamming distance represents the distance between characters, which is not meaningful for continuous variables, so this method is not applicable. However, there is a mix of k-means and k-modes that can be used. This method is called K-Prototype. [2]

The K-Prototype algorithm targets data with mixed attributes. In order to measure mixed data, the distance of numerical attributes is generally calculated using Euclidean distance, for category attributes, we use Hamming distance, if the attribute values are the same, it is 0; if the attribute values are different, it is 1.[3] Finally, adding the two distances together is the total distance. Specifically, the distance between the data and the cluster is:

$$d(X_i, Q_l) = \sum_{j=1}^{p} (x_{ij}^r - q_{lj}^r)^2 + \mu_l \sum_{j=p+1}^{n} \sigma(X_{ij}^c, q_{lj}^c)$$

In this equation, the first sum from j = 1 to j = p are numerical data and from j = p + 1 to j = n are categorical data. r and c are the number of numerical attributes and categorical attributes respectively, and μ_l is the weight of the categorical attributes.

And the objective function of K-prototype is:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_i, Q_l)$$

X_i is the m attributes of sample i, and Q_l is the m attributes of sample l, and y_{il} is the element of the segmentation matrix Y.

With the definition of dissimilarity and prototype, the steps to write down the algorithm are:[4]

1. Input: Number, k, of clusters and the weighting factor
2. Step 1: Randomly select k objects from the data set as the prototype of the initial k clusters and traverse each data point in the data set to calculate the dissimilarity between the data and the k clusters.
3. Step 2: This data is then assigned to the corresponding cluster with the least degree of dissimilarity.
4. Step 3: After each assignment is completed, the prototype of the cluster is updated and then the objective function is calculated, and then the objective function value is compared.
5. Step 4:the cycle is performed until the objective function value does not change.
6. Output: Generate good clusters.

The above is the theoretical principle of K-Prototype clustering. When we finish preprocessing the datathe kmodes.kprototypes library was used to divide the data into two categories according to the

K-Prototype method. Finally, by comparing the predicted classification results with the original classification results, the accuracy of classification can be calculated.

The accuracy found using this method was 89.2%. This method has high accuracy for the data and the method could have helped predict whether the user will checkout if the dependent variable is known. As not much of a result could be derived from the clustering outcome, classifier is trained in the next section to predict classification.

2.5.2 Classification

2.5.3 Model Selection

The ability to precisely classify observations is extremely valuable for various business applications, which usually contain complex relationship among it's independent variables. Likewise, an ideal classification model should also interpret the relationships among it's features in a simple manner besides predicting the response variable accurately. An ideal model for such approach, which handles nonlinear relationships very easily & can produce easily interpretable output, usually referred as Tree based models.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods are often predictive models with high accuracy, stability and ease of interpretation. Decision Trees and the ensemble equivalent, Random Forests, are the most popular algorithms among Tree based models which can deal with both categorical & numeric data & produces a very robust, high-performing model and can even control over-fitting.

2.5.4 Decision Tree

Decision trees are one of the tree based models which is very easy to implement & interpret it's output. In this algorithm the data is split into two or more homogeneous sets based on most significant splitting criteria in input variables. The method of splitting could be Gini Index, Entropy or Chi square. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.[5]

2.5.5 Random Forest

Random Forests are considered one of the most powerful non-linear machine learning algorithms which can handle any type of data and without much preprocessing & produce excellent accuracy, robustness, ease of use & they reduce the chance of over-fitting. A simple decision tree isn't very robust, but a random forest, which runs many decision trees and aggregates their outputs for prediction, often produces a more robust, higher-performing model. Ensemble methods are known to boost tree based models. Random forests use a technique called bagging, which reduces the variance of predictions by combining the result of multiple classifiers modelled on different sub-samples of the same data set.

For given reasons, in this project Random Forest Classifier has been implemented to apply the bagging technique of multiple decision trees & the predicted output has been considered for accuracy of the model by combining the outputs of all decision trees. The data set contained categorical & numerical variables with complex nonlinear relationships among them, therefore Random Forest classifier was implemented to obtain highest accuracy. Gini Indexing is implemented as the splitting criteria for this algorithm.

3 Results

3.1 Results of Question 1

The results obtained from this question is a set of significant observations. *DATE_AND_TIME_OF_PAGE_VIEW* is omitted in the results and not much was to be gained from the findings. The first of which being the trend in device use over time. As Figure 2 shows, the device use looks to change over time. That said, due to the random sampling these trends need to be verified statistically using ANOVA testing to check that the device use is in fact changing instead of just appearing to because of uneven sample sizes for each year.

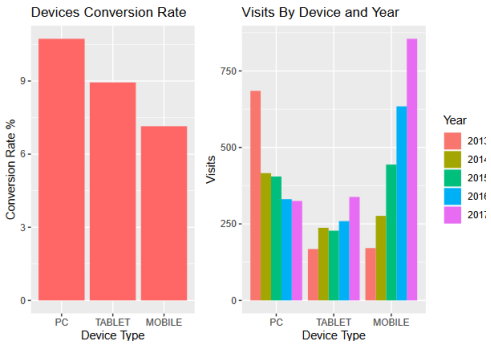

Figure 2: Device use against conversion rate and Device type use by year

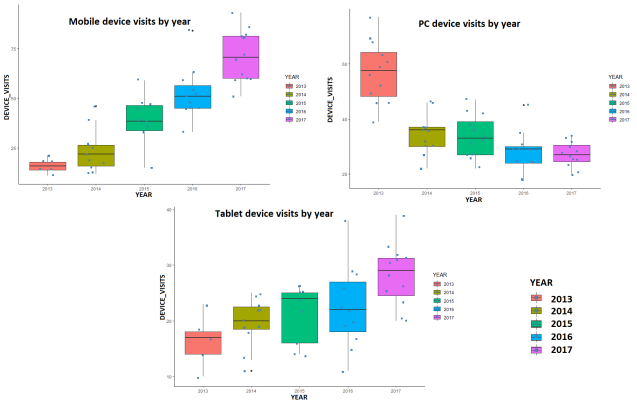

Figure 3: Box plots of device use for each device with time

The box-plots in Figure 3 show more convincingly, that for each of the three devices it would be expected that device use has changed with time. The scales are not identical for each graph as the inter-graph comparison is not important for this Figure, only the differences between boxes in each plot matters.

Hypothesis Studied	P-Value
H_0: Mobile use unchanged with time H_1: Mobile use changing over time	2.035e-13
H_0: PC use unchanged with time H_1: PC use changing over time	1.178e-12
H_0: Tablet use unchanged with time H_1: Tablet use changing over time	0.001817

Figure 4: Each hypothesis test corresponding to each devices use over time

Conducting ANOVA tests to find whether statistically there is any change in use finds, the results in the table (Figure 4). The observed p-values validate, at the 95% level, all three of the trends shown by the previous two Figures. Hence, each device use is changing with time most probably with the trends that are displayed. This is because the ANOVA tests compares the mean values and determines whether they are the same (Null Hypothesis) or different (Alternative Hypothesis).

The second phase of the exploratory analysis considers the trend in merchandise variety of NBrown. This analysis could provide ideas for communication and promotional strategies focusing on external target audiences. These strategies could attract more visitors in a website or increase the sales of promoted products. Prior to more suggestions, a visualisation of the conversion rate with the corresponding visits is required. Similarly, the conversion rate was calculated by dividing the number of checkouts by the number of visits for each merchandise category.

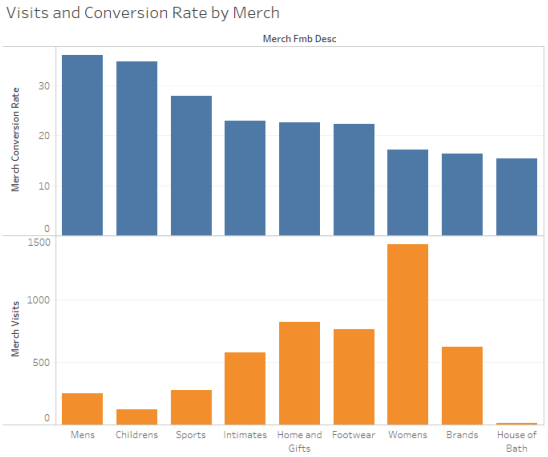

Figure 5: Visits and conversion rate by merch

In Figure 5 is shown that menswear, children's clothing and sportswear are the three merchandise categories with the highest conversion rate, whilst womenswear gather a lot of visits with low conversion rate. Therefore, a proposed tactic would be the advertising funding on the target audiences that consist of males, parents and athletic community. On the other hand, the big female audience indicates that an increase of the sales by increasing the conversion rate would be considerably profitable. This increase could be achieved with discount policies or improvement of the structural platform of websites.

In this question there was potential for bias as it is expected and also found in the data that mobile use would have increased in recent years. However, the statistics overcome this by validating the findings and they are reproducible using the code.

3.2 Results of Question 2

The navigation process of users is analysed on interest of device type and whether or not a web session ends on a checkout. Combining the two variables, four test cases are created based on the aforementioned links between Home, Navigation, Product View, Search Results and View Bag pages. At the end of each session, it is added a row indicating the exit of the user in order to retrieve better insight from the visualisation of navigation experience. In summary, the device type does not differentiate the process and the main interest is focused on the pages before the exit of a session that ended on abandoning a bag or an order. The results can be shown in Figures below.

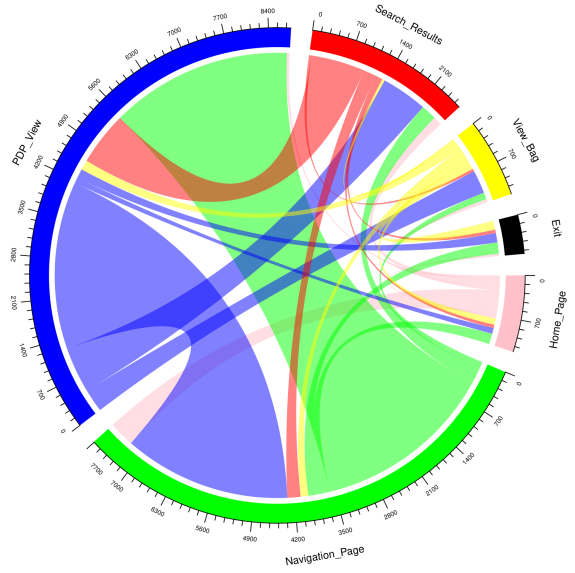

Figure 6: Chord diagram of abandon bags and orders

Last Page Visited	Next Page	Number of occurrences
Home_Page	Exit	41
Navigation_Page	Exit	198
PDP_View	Exit	161
Search Results	Exit	62
View_Bag	Exit	155

Figure 7: Table of links between last page and exit

Initially, in Figure 6 is obvious that most transitions are achieved between Product View and Navigation page. This means that these pages are playing a crucial role in the navigation experience of the user and need to be in the main interest of web page developers. Furthermore, combining the chord diagram and the table in Figure 7, most users quit their session after the aforementioned pages and the View Bag page. According to these pages, some factors that may lead on abandoning the basket could be summarised as follows:

1. Defective structure of Navigation page and difficulties of exploring a variety of products
2. Deficient amount of product information in Product View page
3. Complex checkout process in View Bag page

To conclude, tracking the navigation experience of users can provide valuable information about the faulty operation and structure of the system.

3.3 Results of Question 3

Figure 8 shows a decision tree out of the created random forest model. It plastically visualizes how the model works. Features which are capable of creating a higher node purity when used for splitting the data are used in higher levels during the process of tree building to split. Purity is given by the Gini coefficient, which also can be used to derive the feature importance weighted by the probability of reaching that node. The random forest built during this project was created using 10 trees and restricting the maximal depth to 5. Given a test split of 70:30, it predicts 125 out of 150 "no checkout" samples correctly, while 137 times a checkout is predicted accurately. This yields an accuracy of 87%, a precision of 85%, while recall is at 91% and the F1 score at 0.88, so the predictive performance of the forest can be considered good.

The feature importance extracted out of the forest is visualized in Figure 9 and indicates, which features are having the meaning for predicting whether a checkout is performed or not. Herein, the total number of visited pages is the most important feature with 40% importance. The classification is also highly impacted with 30.5% by the total time spent browsing. Whether the user is an account holder or not and which month of the year the page is visited contributes with 16%, respectively 11%. Hence, these features are the basis for proposing improvement measures for increasing the number of checkouts. The remaining two features "Device type" and "New Visitor Index" do contribute very little to prediction quality.

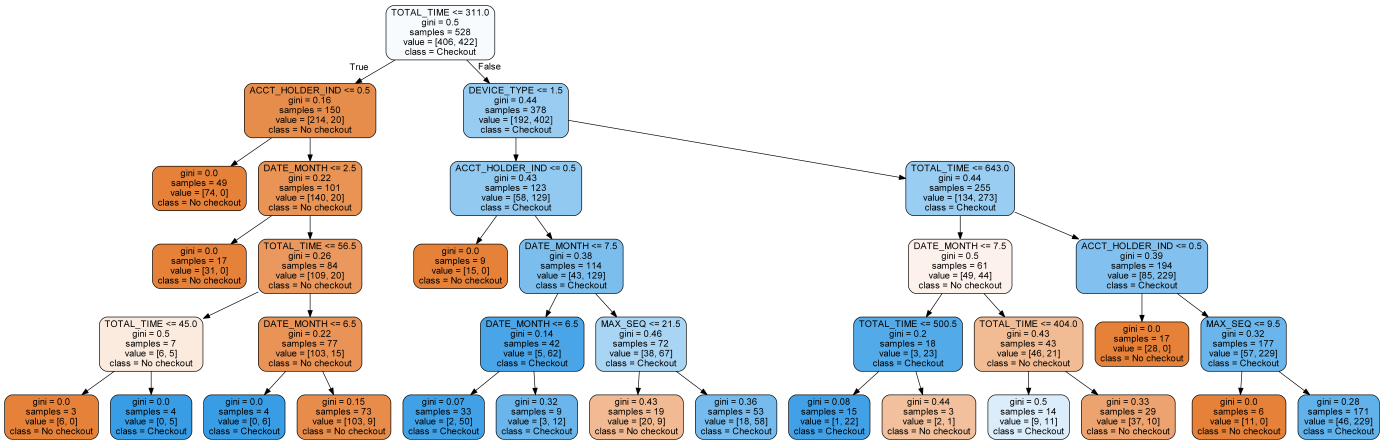

Figure 8: Decision Tree with 6 features

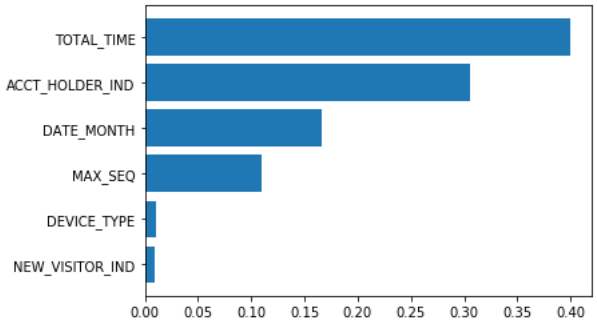

Figure 9: Feature importance

However, random forests are hard to interpret due to the fact that feature distribution among trees is performed randomly, causing each tree to look different. For better interpretation, a single decision tree is fitted with the 4 important features, identified in the previous section. The result can be seen in Fig 1. From the diagram, different dependencies can be derived. Its root node shows, that a checkout is very unlikely, if the total time of visit is less than 5 min. Also registered customers browsing the sites for more than 5 minutes via mobile in the second years half are likely to checkout, if having more than 21 hops. To increase the probability of a checkout for a new visit, one measure would be to keep people on the page once initially visited, by displaying style recommendations based on last purchases on landing page to get people browsing, or by displaying banners for short time promotion codes.

For registered mobile users staying longer than the critical 5 minutes threshold, checkout probability is very high in the first years half in general. To increase the number of visits within this target group, getting users on the page by sending push notifications on mobiles may be appropriate. To then increase the turnover rate within the group, advertising placed prominently on the website, targeted on registered customers, with special offers or limited products could be set up.

Registered users on all other devices tend to make a purchase, if staying longer than than 11 minutes and doing more than 9 hops, or in the first half of the year, when staying less than 10 minutes.

It is important to state, that for the above proposals, 2 assumptions are made. The first one is that duration and sequence of visit are causal for checkouts, and not only correlated. This means, that by increasing the time that users spend on the sites, the likeliness of checking out increases. The second assumption is, that the it is possible to checkout as a guest, so no registration is mandatory. This is for the variable of user registration to hold the assumption of independence, as it else could not be used as explanatory variable.

4 Conclusions

Throughout the project, three research questions have been defined and were investigated: Using the approach of exploratory data analysis, as well as classification, the most important influential factors on the user behaviour for checking out have been exposed. According to these, fully registered visitors who spend longer than 10 minutes browsing in the first half of the year are most likely to check out. These findings were used to make proposals for increasing the conversion rate.

The way how visitors are navigating the pages has been analyzed and was depicted using appropriate graphs and a model for being able to lively calculate the probability of a user checking out has been built, which is of tremendous value, based on which exemplary proposals were made to increase this likeliness.

However certain assumptions on causality had to be made, so that recommendations can be drawn. Further than that, for modelling, only simple random undersampling was performed. Combining this with further techniques such as SMOTE may yield even better predicive performance as more samples of the majority class could be considered.

A first step towards considering the web flow data for check out modelling has been made, which provides as scope for further work, to analyse it in greater detail, by using feature engineering methods based on the already performed exploration, to synthesize the important sequential features influencing the checkout behaviour and integrate them into our model.

References

[1] N Brown Group plc. *About us*. 2019. URL: https://www.nbrown.co.uk/about-us.

[2] 沙漠之狐. *Introduction to clustering algorithms k-means, k-modes, and k-prototype*. 2019. URL: http://blog.sina.com.cn/s/blog_b92ab6cb0102veeg.html.

[3] 奥特曼打小怪兽. *Select clustering algorithms based on variable attributes (K-means, Kmodes, K-prototype*. 2018. URL: https://www.jianshu.com/p/c9dcc52b85d4.

[4] 不矜不伐的小学生. *k-prototype algorithm analysis*. 2016. URL: https://www.cnblogs.com/zhangruilin/p/5769795.html.

[5] Analytics Vidhya. *A Complete Tutorial on Tree Based Modeling from Scratch*. 2019. URL: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/.