

**Predictive Modelling for *TechScape***

**Machine Learning | MDSAA 2021/22**



**Group 60**António Fonseca - r2018…  
João Carvalho - r2018…  
Lucas Lopes - r2018…  
 Mariana Ferreira - r20181071  
Tiago Oom - r20181077

**Professors:** Roberto Henriques, Carina Albuquerque and Lara Oliveira

**Index**

[**1.** **Introduction** 2](#_Toc89547072)

[**2.** **Data preparation** 2](#_Toc89547073)

[**3.** **Dataset exploration** 2](#_Toc89547074)

[*3.1.* *Descriptive statistics* 2](#_Toc89547075)

[*3.2.* *Coherence checking* 3](#_Toc89547076)

[*3.3.* *Missing values* 3](#_Toc89547077)

[*3.4.* *Outliers* 3](#_Toc89547078)

[**4.** **Feature engineering** 3](#_Toc89547079)

[**5.** **Feature selection** 3](#_Toc89547080)

[**6.** **Model training** 3](#_Toc89547081)

[*6.1.* *Data partition* 3](#_Toc89547082)

[*6.2.* *Data preprocessing* 3](#_Toc89547083)

[*6.3.* *Models* 3](#_Toc89547084)

[**7. Model deployment** 3](#_Toc89547085)

[**8. Discussion** 3](#_Toc89547086)

[**9. Conclusions** 3](#_Toc89547087)

[**10. References** 3](#_Toc89547088)

[**11. Appendices** 3](#_Toc89547089)

**Index**

[**1.** **Introduction** 2](#_Toc89547054)

[**2.** **Data preparation** 2](#_Toc89547055)

[**3.** **Dataset exploration** 3](#_Toc89547056)

[*3.1.* *Descriptive statistics* 3](#_Toc89547057)

[*3.2.* *Coherence checking* 3](#_Toc89547058)

[*3.3.* *Missing values* 3](#_Toc89547059)

[*3.4.* *Outliers* 3](#_Toc89547060)

[**4.** **Feature engineering** 3](#_Toc89547061)

[**5.** **Feature selection** 3](#_Toc89547062)

[**6.** **Model training** 3](#_Toc89547063)

[*6.1.* *Data partition* 3](#_Toc89547064)

[*6.2.* *Data preprocessing* 3](#_Toc89547065)

[*6.3.* *Models* 3](#_Toc89547066)

[**7. Model deployment** 3](#_Toc89547067)

[**8. Discussion** 3](#_Toc89547068)

[**9. Conclusions** 3](#_Toc89547069)

[**10. References** 3](#_Toc89547070)

[**11. Appendices** 3](#_Toc89547071)

## **Introduction**

(Introduzir o tema, objetivos e metodologia)

## **Data preparation**

(Dataset inicial e alteração nos tipos de variáveis) --vski

## **Dataset exploration**

## *Descriptive statistics*

(Análise das estatísticas descritivas e dos gráficos) -- Lucas

## *Coherence checking*

(Explicar o raciocínio) -- Lucas

## *Missing values*

(Explicar que não há) --Lucas

## *Outliers*

(Metodologia adotada para lidar com outliers) -- antonio

## **Feature engineering**

Besides the original variables of the dataset, we created new … variables by combining or discretizing the initial ones to increase the number of possible independent variables that influence our target. This feature engineering process also included some adjustments of incoherent values to allow correct calculations, namely on the new variables Total\_Pages, Total\_Duration, AvgTimeOnAccountMng, AvgTimeOnFAQ and AvgTimeOnProduct.

These new variables and the description of their engineering process are presented below in table 2.

|  |  |
| --- | --- |
| New variable | Description |
| Apple | 1 if OS equals MacOSX or iOS, 0 otherwise |
| Windows | 1 if OS equals Windows, 0 otherwise |
| Android | 1 if OS equals Android, 0 otherwise |
| Returner | 1 if Type\_Of\_Visitor equals Returner, 0 otherwise |
| New\_Access | 1 if Type\_Of\_Visitor equals New\_Access, 0 otherwise |
| Total\_Duration | AccountMng\_Duration + FAQ\_Duration + Product\_Duration |
| Total\_Pages | AccountMng\_Pages + FAQ\_Pages + Product\_Pages |
| PropTime\_AccountMng | AccountMng\_Duration/Total\_Duration |
| PropTime\_FAQ | FAQ\_Duration/Total\_Duration |
| PropTime\_Product | Product\_Duration/Total\_Duration |
| PropPages\_AccountMng | AccountMng\_Pages/Total\_Pages |
| PropPages\_FAQ | FAQ\_Pages/Total\_Pages |
| PropPages\_Product | Product\_Pages/Total\_Pages |
| AvgTimePerPage | Total\_Duration/Total\_Pages |
| AvgTimeOnAccountMng | AccountMng\_Duration/AccountMng\_Pages |
| AvgTimeOnFAQ | FAQ\_Duration/FAQ\_Pages |
| AvgTimeOnProduct | Product\_Duration/Product\_Pages |
| Month | Month extraction from Date |
| Covid | 1 if Month is greater than 3, 0 otherwise |
| Browser | One new variable per each Browser type |
| Country | One new variable per each Country |

***Table 2*** *– New variables.*

## **Feature selection**

(Processo de decidir as variáveis, explicar o raciocínio do k-folds) –johnny

**5.1 Correlation Matrix**

For the feature selection section of the project, we first decided do check for any univariate variables since it would be pointless to keep one in the model as they wouldn´t provide any relevant information. There was none.

Then we resorted to the Pearson matrix of correlations to search for any pair/subset of variables which were too correlated with each other since keeping these variables in a model would harm its performance. Note that this step was only performed on numerical variables as it isn´t adequate for categorical variables.



The following table shows the most problematic pairs of variables that will not be included in the model simultaneously. However, the correlation matrix was consulted and taken into account whenever trying new combinations of variables for our models.

|  |
| --- |
| Problematic Variables |
| Product\_Pages | Product\_Duration | 0.86 |
| Product\_Pages | Total\_Pages | 0.85 |
| Product\_Duration | Total\_Pages | 0.86 |
| Product\_Duration | Total\_Duration | 0.99 |
| BounceRate | Exit\_Rate | 0.91 |
| PropTime\_AccountMng | PropPages\_Product | -0.76 |
| AvgTimePerPage | AvgTimeOnProduct | 0.98 |

**5.2 Stratified k-Folds for numerical data**

Next, we proceeded to a more objective process of feature selection. More concretely we performed 7 feature selection algorithms: *Lasso Regression; Decision Tree (Gini); Decision tree (Entropy); Regressive Feature Elimination; Backward Regression; Forward Selection; Select From Model*. Note that this algorithms were also only performed on numerical variables as they aren´t suitable for categorical variables.

For those, we decided to use stratified 10-folds to split our data into train and validation and performed each one of those using 10 different train and validation datasets. This way we were able to avoid the potential bias caused by using only one regular split of the data. In the end, we were able to visualize which variables were constantly identified as relevant for our model. The results overview is displayed below.

Uma imagem com mesa

Descrição gerada automaticamente

**5.3 Stratified k-Folds for categorical data**

For the categorical features we performed the exact same analysis but with only one algorithm, the *Chi-Square test for categorical data*. The results overview is displayed below.

Uma imagem com mesa

Descrição gerada automaticamente

At the end of this feature selection section, we were able to identify which variables are plausible candidates to be included in the final model. However, the final results are just a guideline as we try new combinations of variables and evaluate their performance.

## **Model training**

## *Data partition*

(Explicar o objetivo do k-folds) --mariana

## *Data preprocessing*

(Tratamento do train depois da partição) --johnny

## *Models*

(Desenvolvimento dos modelos) --vski

## **7. Model deployment**

(Implementação do melhor) --vski

## **8. Discussion**

(Discussão dos resultados – comparação de vários modelos?)

## **9. Conclusions**

(Contextualização dos resultados no problema e limitações)

## **10. References**

## **11. Appendices**