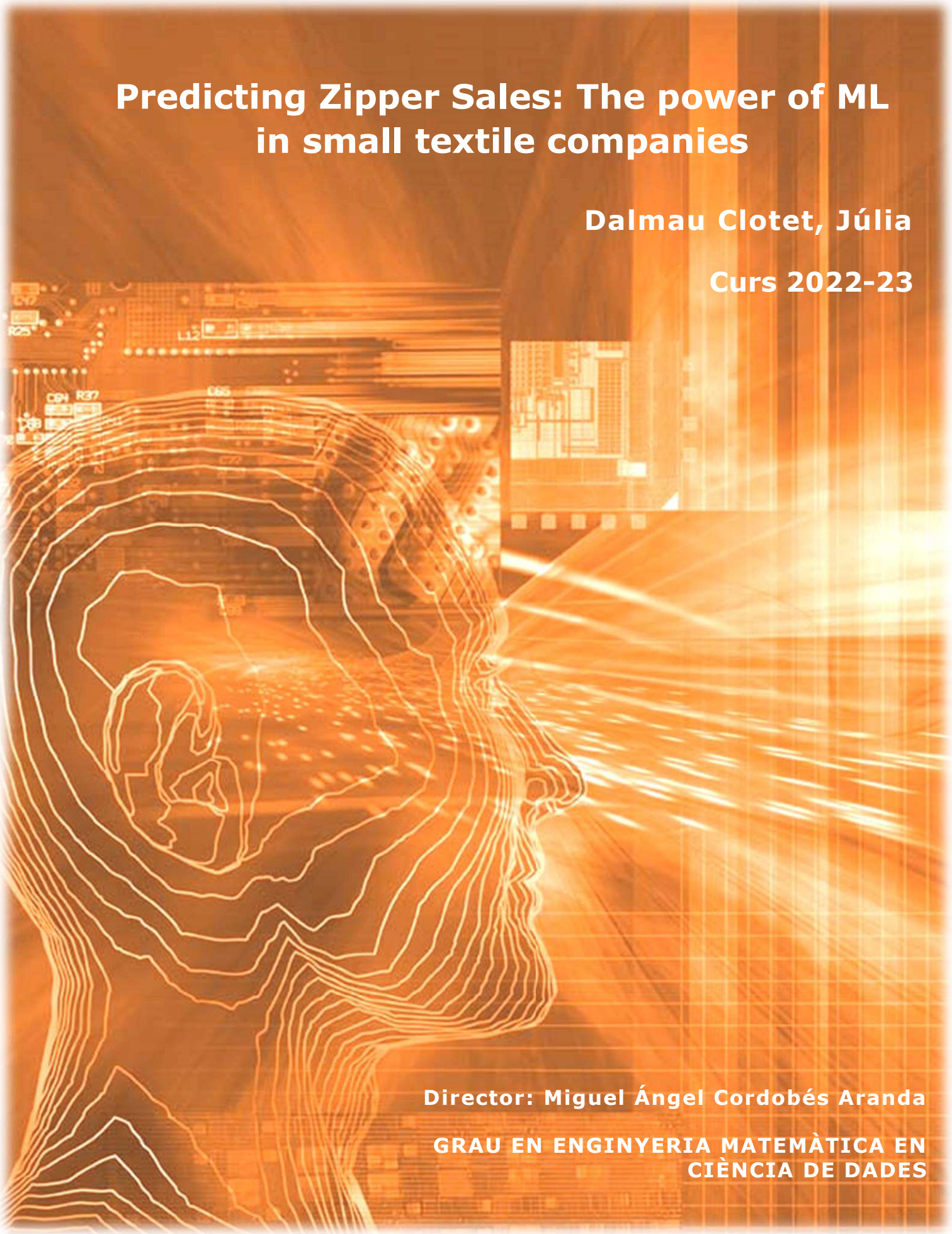
Predicting Zipper Sales: The power of ML in small textile companies

Dalmau Clotet, Júlia

Curs 2022-23



Director: Miguel Ángel Cordobés Aranda

GRAU EN ENGINYERIA MATEMÀTICA EN CIÈNCIA DE DADES

upf. Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Treball de Fi de Grau

Acknowledgements

Primarily, I want to convey my heartfelt gratitude to Kreband for the opportunity to collaborate with them on a Machine Learning project. I am grateful for the opportunity to work on a real-world problem and to be a part of this intriguing initiative. Your trust and confidence in me will not be forgotten.

Additionally, I would like to extend my deepest appreciation to my thesis tutor, Miguel Ángel Cordobés. His patience, dedication, and guidance have been instrumental throughout every step of this project. I am incredibly grateful for his constant support and encouragement, which have helped me to achieve my goals.

Also, I want to say thank you to my family for the unconditional support and patience that they provide me during all the thesis process and all the four years of the bachelor's degree. I could not have made it this far without you.

Finally, I would like to express sincere appreciation to one of my home friends for all the help and patience along this process, and to conclude, I would like to make a special mention of my Erasmus friends. Those who have become my new family over the last six months, who gave me their time to advise, listen to, and assist me in the project as well as daily. Your love, energy, knowledge, and patience are reflected along this thesis, as well as they will be forever with me. Thank you from the bottom of my heart!

Abstract

The textile industry has begun to embrace algorithm and Machine Learning (ML) predictions over the past decade to improve their sales performance and overall efficiency. In this project, ML models are applied in a zipper dataset to try to find an accurate prediction about future sales trends for a small company called Kreband. Furthermore, an analysis of the zippers features is applied to see which is the impact in the ML model and how the company can reduce costs and gain efficiency to compete against the multinational companies.

Resum

Durant la darrera dècada, la indústria tèxtil ha començat a descobrir els algorismes i les prediccions de l'aprenentatge automàtic per millorar el seu rendiment. En aquest projecte, s'apliquen diferents models d'aprenentatge automàtic a un conjunt de dades de per tal de trobar la predicció més precisa possible sobre les vendes de cremalleres per l'empresa Kreband. A més, s'usa un anàlisi de les característiques de cada tipus de cremalleres per veure quin és l'impacte en el model i així ajudar a l'empresa a reduir les despeses i guanyar eficiència per poder competir davant les grans companyies multinacionals.

Resumen

Durante la última década, la industria textil ha empezado a descubrir los algoritmos y predicciones del aprendizaje automático para mejorar su rendimiento. En este proyecto, se aplican diferentes modelos de aprendizaje automático a un conjunto de datos para encontrar la predicción más precisa sobre las ventas de cremalleras por la empresa Kreband. Además, se usa un análisis de las características de cada tipo de cremalleras para ver cuál es el impacto en el modelo y así ayudar a la empresa a reducir los gastos y ganar eficiencia para poder competir frente a las grandes compañías multinacionales.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

- Machine Learning (ML)
- Supervised Learning (SL)
- Support Vector Machines (SVM)
- Extreme Gradient Boosting (XGBOOST)
- Random Forest (RF)
- Root Mean Squared Error (RMSE)
- Receiver Operating Characteristic (ROC)
- R-squared (R2)
- Mean absolute error (MAE)
- Machine Learning Operations (MLOPS)
- Exploratory Data Analysis (EDA)
- Shapley Additive Explanations (SHAP)
- Deep Learning (DL)
- Recurrent Neuronal Networks (RNN)

1. INTRODUCTION

1.1 Background

Since the last decade, textile industry has begun to use ML algorithms and predictions to improve the sales and predict the next fashion trends. The multinational companies like Inditex are already using these new techniques to forecast demand for products and optimize the production. However, small textile companies with more limited financial resources have just started to see the benefits of ML and opportunity to try to compete against the big multinationals.

This is the case of the familiar company called Kreband. Kreband is a company situated near Barcelona, which manufactures and distributes reinforcement tapes for footwear and leather industries, as well as label material for textile and industrial sectors [1]. Kreband is trying to compete against the huge companies since long time ago, which it is usually innovating in new areas of the textile world. This is how Kreband ended up in the zipper's world, fashion trends and ML algorithms.

The problem that the company has is that manufactures the zippers in China and spends a lot of resources and money in transportation. Therefore, when a customer's order arrives, the corporation must ship the zippers from China to Spain as soon as feasible. Today, the fastest mode of transportation is air travel, which is fast but, as we all know, can be quite expensive. As a result, Kreband wishes to spend less money on transportation by using another mode of transportation, such as the sea. To do so, they want to know which zippers will be sold in the upcoming sessions based on previous fashion trends to anticipate client purchases and have the zippers ready for customer requests.

1.2 Objectives

The main objective of this project is to analyse the zippers dataset, extract and find a good model sales forecast that can help the company to reduce their costs and improve efficiency with their clients. Therefore, the objectives of this work are:

1. Familiarize with the Machine Learning Operations (MLOPS) process to develop a ML project.
2. Analyse the dataset using visualization techniques.

3. Develop different ML techniques to find the best one for our data.
4. Analyse the different ML algorithms and techniques offer more precision results and describe the advantages and limitations of the models.
5. Analyse and visualize the impact of the zippers feature in the best ML model.
6. Create a webapp that has the most accurate ML model with the intention to display the results and do forecasts.

2. USE CASE

The use case of this project is to apply Artificial Intelligence, in particular ML, to a manufacture company to improve its logistics. In particular, the objective is to predict which zippers will be sold in upcoming seasons to optimize production, reduce transportation costs, and improve customer satisfaction.

With this objective, the first step is to do a data analysis to get information about the client orders of the past and to evaluate if there is a fashion trends in our data. After, different ML algorithms will be applied to the data to see which one can estimate demand for different kind of zipper in forthcoming seasons. Later, an interface will be created to display the results to the company and to show them the opportunity to see which zippers will be sold or not.

3. STATE OF THE ART

In this project, ML models are used to produce predictions or choices. ML models are data-trained algorithms that detect patterns or relationships in the data and then, utilize this information to generate predictions or classify unseen data based on its properties. Usually, the ML models are divided into: Supervised Learning (SL) and Unsupervised Learning Models.

The main difference between these two types is that SL models are trained using samples with a label to predict, while the Unsupervised Learning models train on unlabelled data to try to find patterns in the results groups or clusters [4].

In our case, as we have a zipper labelled dataset with features, and the final part is to predict the total sales, we will develop only SL models. For finding patterns in our data, we will develop two different techniques to see which one performs more efficiently. These two types are[1]:

- **Classification**

  The classifiers assign test data into specific categories. We will develop different types of classifiers:

  - Extreme Gradient Boosting (XGBOOST)
  - Random Forest (RF) Classifier
  - Support Vector Machines (SVM)
  - Gradient Boosting
  - Light GBM
  - CatBoost Classifier

- **Regression**

The regression is used to see the relation between dependent and independent variables and try to predict numerical values based on different data points. The regressors used are:

  - Linear Regression
  - CatBoost Regressor

[1] See Appendix 10.1 for the explanation of the algorithms.

4. IMPLEMENTATION AND DESIGN

In this section, we describe the whole end-to-end process focused on this ML project.

4.1 MLOPS framework

To develop this project, the MLOPS framework was used to consider all aspects and activities done in a ML project. The process followed for the project is:

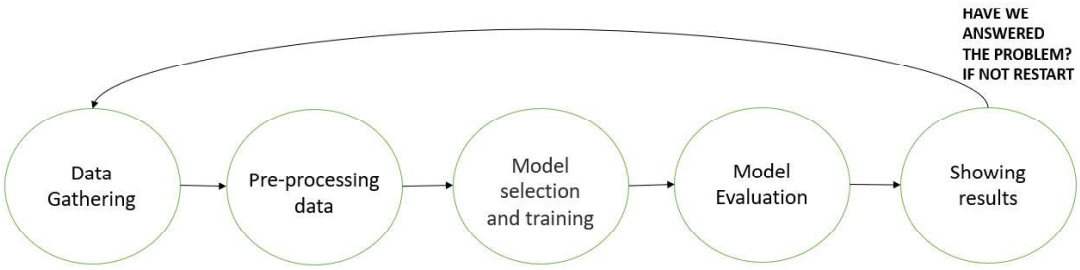


FIGURE 2: MLOPS OUTLINE

1. **Data Gathering**

   The company provided the data. The dataset has the zipper features and the sales of the last years.

2. **Data Pre-processing**

   After gathering the data, it was cleaned, formatted, and transformed to a form that can be used for the ML models. During this process, tasks like handling missing values and removing duplicates have been applied to get more consistency to our dataset.

3. **Model selection and training**

   In this part, different types of ML algorithms were applied to find the better algorithm that fits to our data. The first step was to apply a binary classification differentiating the zippers that were sold and the zippers that were not sold. The second technique applied was trying the regressors models. After applying these techniques, as the results were not as we expected, another approach was performed. This new approach was to do a

multiclass classification to determine in which quartile the new zipper will be sold according to the past sales.

Here, we can find an outline about the model selection process followed:

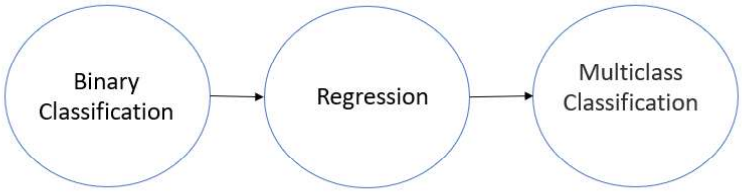


FIGURE 3: MODEL SELECTION OUTLINE

4. **Model Evaluation**

- After training the different models, an evaluation was performed to determine the performance. Measures like accuracy, precision, recall, F1 score, Root Mean Squared Error (RMSE), Mean absolute error (MAE), R-squared (R2) were studied. Also, an explain ability of the ML algorithms using the Shapley Additive Explanations (SHAP) values was done to get a better understanding of the models.

5. **Showing results**

The final step is to create an interface which the company can use to see the results of the best ML algorithm and can use this interface to predict the zippers that can be sold.

4.2 The data

A dataset with 4699 samples and 64 features was provided by the company to complete this project. The dataset includes the features and sales of footwear zippers over the last four years for the company, starting on January 2018 until May 2022.

The first ten attributes of our data explain how a zipper is composed, and the other 54 attributes correspond the number of zippers sold every month since January 2018. Before starting with the data cleansing, it is important to know which type of data are the attributes.

The zippers are composed by ten features:

- **Length:** the length is the magnitude of the zipper. The data has 69 different types of length, and they are composed with three figures, which can be numbers and capital letters.
- **Colour:** is the colour of the zipper. In our case, 180 different types of colours can be find, and they are named using numbers and capital letters.
- **Family:** We have fifteen different families in the dataset which are composed by three elements:
    - Type**:** it consists of the material that the zipper is created. There are sixteen different materials, and they are named using capital letters.
    - Number**:** We find seven different types of numbers which brings to us information about the utility of the zipper. They are labelled using numbers and the letter A. The numbers start in the 3 and end in the 9 plus the letter A. In our case, it will be always the number seven, because we are working on zippers that will be used on the footwear world.
    - Form**:** it brings information about the type of ending of the chain. There are ten different types, and they can be labelled using capital letters and numbers.
- **Teeth:** the teeth are the elements of the zipper that mesh or interact with each other when passed through the slider [2]. This feature indicates the type of material which they can be. We can find seventeen types of teeth, but in our dataset, we only have three types. They are labelled using capital letters.
- **Stoppers:** stoppers are two mechanisms attached to the top of a zipper to keep the slider from falling off the chain [3]. We can differentiate between:
    - Top Stopers: This one avoids the slider to go out of chain. There are thirty-one different types of top stopers.
    - Bottom Stopers: avoids the slider to go out of chain. There are twenty-three different types of bottom stoppers.

  For simplicity, we added both stopers only in one variable called stoppers. In total, we have 26 types of them, and they are named using capital letters.
- **Sliders:** the sliders join or split the elements when the zipper is opened or closed [2]. There are 93 different types of sliders which are named using some capital letters and numbers, for example: 01F, 16F,…
- **Label:** the label allows the company to determine if a client preferred their own label or the label from Kreband. We only have 6 different types of labels.
- **Code:** this feature is produced by combining all the previous attributes.

- **Description:** attribute that contains a brief explanation of the zipper.
- **Code without label:** all the prior elements were put together to generate this feature.

Here we can find a figure which show us some parts of the zipper:

FIGURE 4: ZIPPER FEATURES

All the features are type object when the data set is read in the Jupyter Notebook, and the other 54 attributes are integer numbers. It is important that before importing the dataset all the values in the Excel file are as type general to do not have any controversial with the commas and the points when we use the data.

4.2.1 Data Cleansing

To do the Exploratory Data Analysis (EDA) and the modelling part of the project, a data cleansing was applied to our initial data. The modifications done in our data were:

1. All the missing values of the data were deleted in order to give more consistent to our data.
2. The columns: *'Codi'*, *'Codi sense etiqueta'*, *'Descripció'* were deleted. They were eliminated since they provided more information that we could obtain from the other features.

3. The quantity of zippers of every month since 2018, were grouped by quarters. This process was made to know the quantity of zippers that was sold in each quarter, because the company wants the prediction for each quarter.
4. The type of data of the zippers sales grouped by quarters was changed from a float type to integer tap to work better in the future steps.
5. Before applying the ML algorithms, an encoding was used to convert the categorical data to numerical data. This process was applied to facilitate the EDA and to be able to apply our data to a ML models that only can support numerical data analysis.

In the following figures, we can see the original dataset and the dataset with the modifications that were applied:

	Codi	Descripció	Familia	Stopers	Sliders	Teeth	Color	Label	Llargada	Codi sense etiqueta	Quantitat: Gen/2018	Quantitat: Feb/2018	Quantitat: Mar/2018	Quantitat: Abr/2018	Qua Ma
0	B7CAA01F01335012	ZIP PLTD N7 CE AL.AL 01(71).NF C-335+01 SL 12 CM.	B7C	AA	01F	1	335	0	12	B7CAA01F1335012	0.0	0.0	0.0	0.0	
1	B7CAA01F01400012	ZIP-PLTD N7 CE AL.AL 01(71).NF C-400+01 SLV 12 C	B7C	AA	01F	1	400	0	12	B7CAA01F1400012	0.0	0.0	0.0	0.0	

FIGURE 5: ORIGINAL DATA

	Familia	Stopers	Sliders	Teeth	Color	Label	Llargada	1r trimestre 2018	2n trimestre 2018	3r trimestre 2018	4t trimestre 2018	1r trimestre 2019	2n trimestre 2019	3r trimestre 2019	4t trimestre 2019	1r trimestre 2020
0	B7C	AA	01F	1	335	0	12	0	110	0	0	0	0	0	0	0
1	B7C	AA	01F	1	400	0	12	0	140	0	0	0	0	0	0	0
2	B7C	AA	01F	1	400	0	16	0	0	0	100	0	0	0	0	0

FIGURE 6: DATA NOT LABELLED

	Familia	Stopers	Sliders	Teeth	Color	Label	Llargada	1r trimestre 2018	2n trimestre 2018	3r trimestre 2018	4t trimestre 2018	1r trimestre 2019	2n trimestre 2019	3r trimestre 2019	4t trimestre 2019	1r trimestre 2020
0	0	0	0	0	0	0	0	0	110	0	0	0	0	0	0	0
1	0	0	0	0	1	0	0	0	140	0	0	0	0	0	0	0
2	0	0	0	0	1	0	1	0	0	0	100	0	0	0	0	0

FIGURE 7: DATA LABELLED

4.3 EDA

To have an initial perspective about how the data is distributed Python and Tableau were used. These tools let us observe important aspects about the zippers sales of the last years and how is our data is distributed.

Here we can find some graphics:

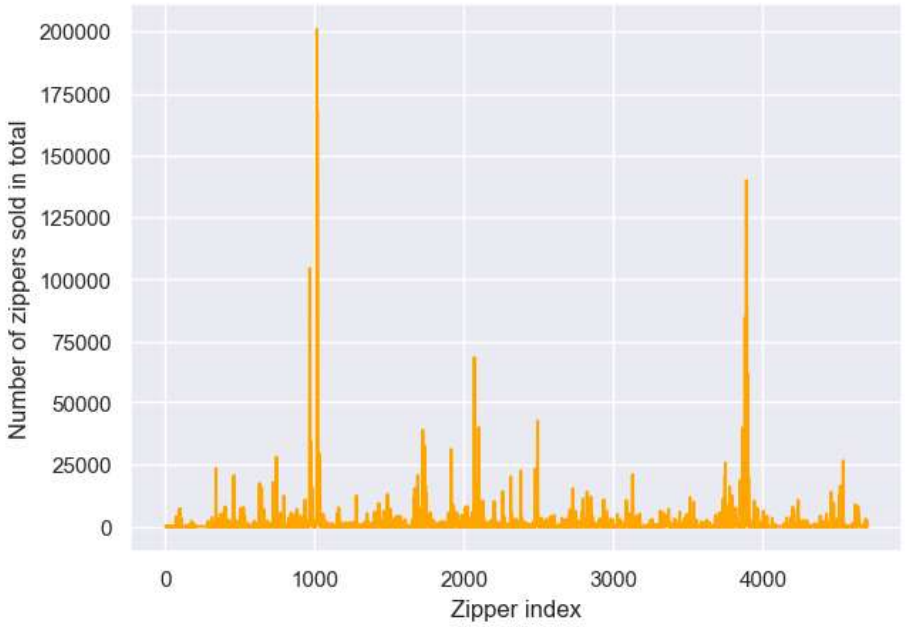


FIGURE 8: QUANTITY OF ZIPPERS SOLD IN TOTAL

In this plot, the x-axis represents the zipper index of our dataset and the y-axis, represents the number of zippers sold in total from the last years. From this graphic, we can appreciate that there is a significance difference between the quantity of zippers sold, since we have some zippers that were almost not sold or directly not sold, and other that the quantity of zippers sold is huge.

The next graphic shows the number of zippers sold each month since 2018. We can see that there is an increase in the total number of zippers sold at the beginning of the year and a drop in the total number of zippers sold throughout the summer months. Another rise can be seen at the end of summer and mid-October, followed by a reduction throughout the rest of fall until winter arrives. Furthermore, if we look at the year 2020, we can see that it is not as we previously stated; this might be due to COVID-19.

Here we can find the plot:

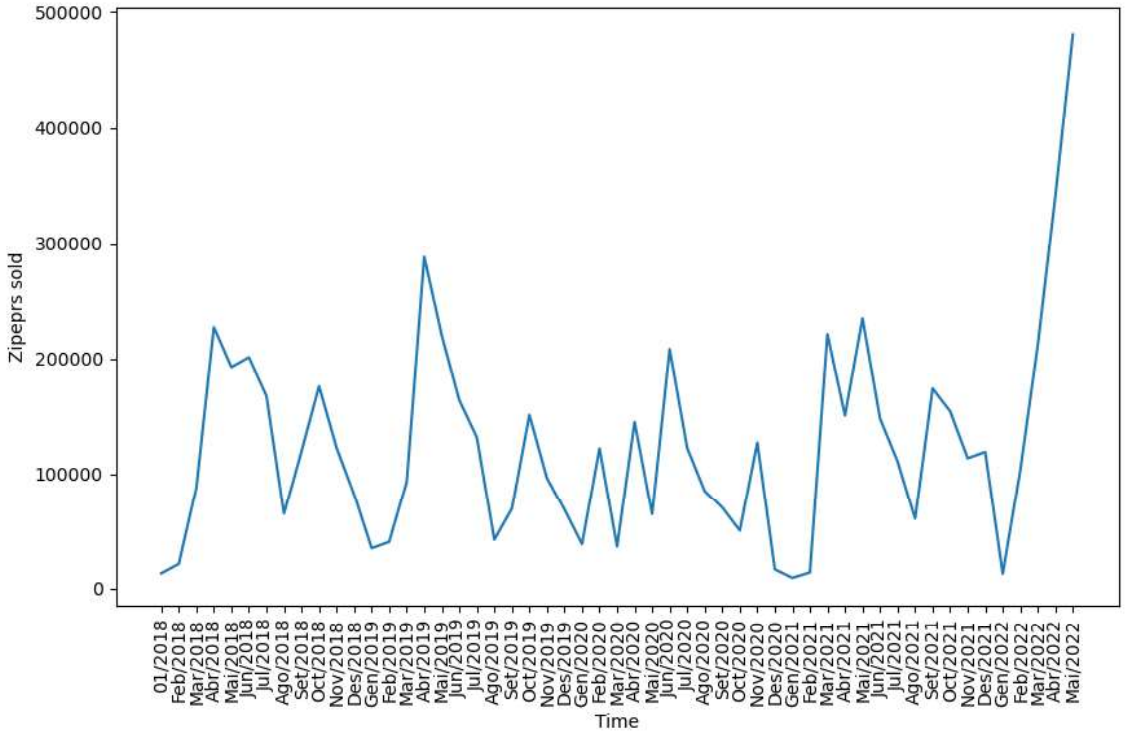


FIGURE 9: QUANTITY OF ZIPPERS SOLD FOR EACH MONTH

Moreover, to continue exploring the data, different plots according to the top five best-selling varieties of each feature over the years were plotted. In the next graphics, we can see the results:

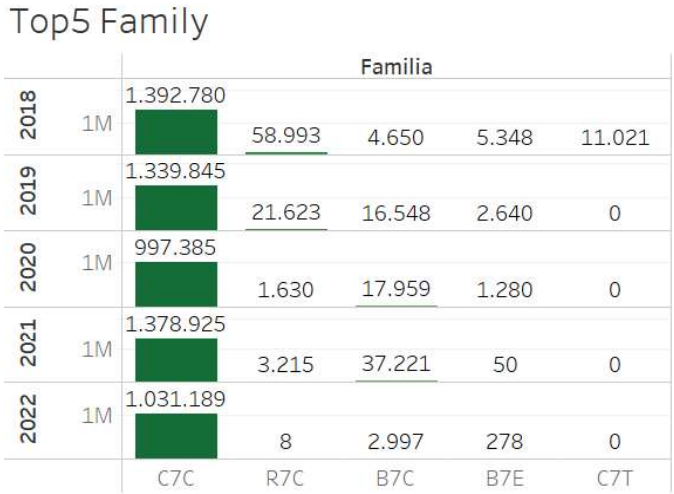


FIGURE 10: TOP 5 FAMILY OVER THE YEARS

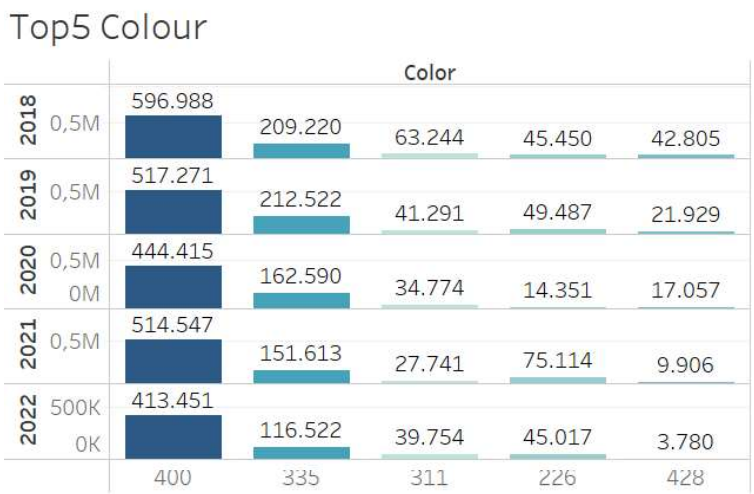


FIGURE 11: TOP 5 COLOUR OVER THE YEARS

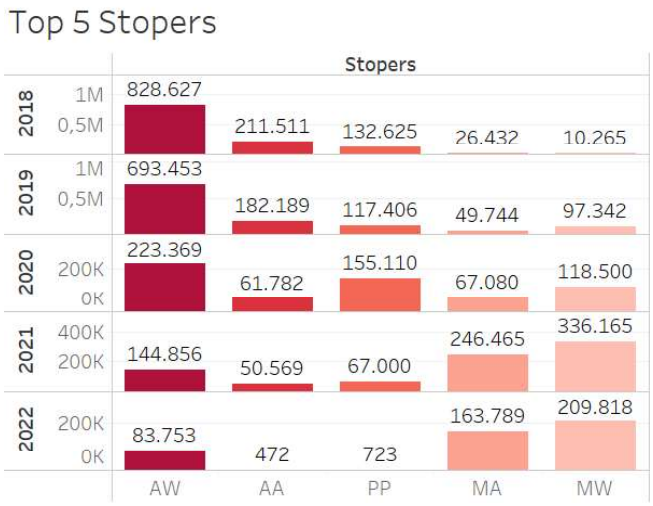


FIGURE 12:TOP 5 STOPERS OVER THE YEARS

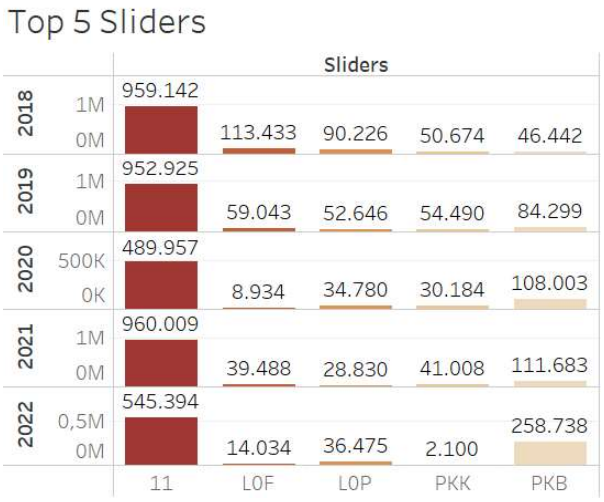


FIGURE 13: TOP 5 SLIDERS OVER THE YEARS

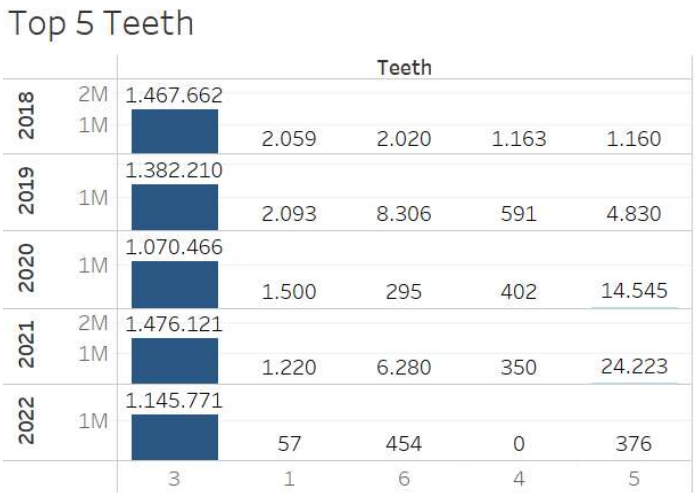


FIGURE 15:TOP 5 TEETH OVER THE YEARS

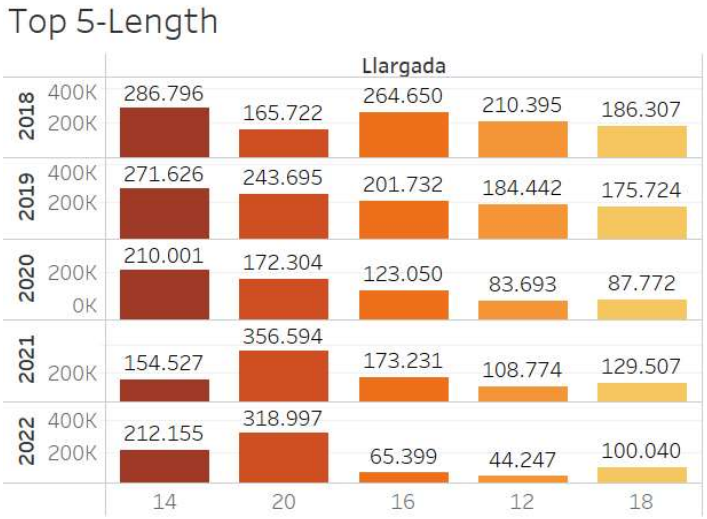


FIGURE 14: TOP 5 LENGTH OVER THE YEARS

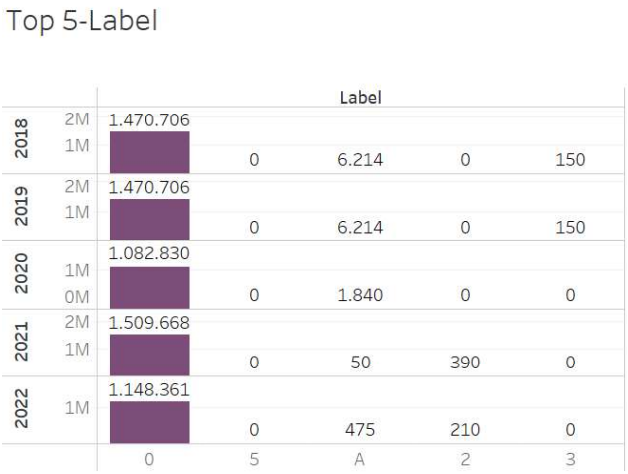


FIGURE 16:TOP 5 LABEL OVER THE YEARS

The x-axis shows the top five zipper feature kinds by sales over the last several years, while the y-axis shows the amount sold of each zipper feature type per year. These graphs show the most prevalent types of each attribute across time, and we can observe that the top characteristics do not change significantly over time. However, we can see that the number of each kind drops in 2020, because of COVID-19, and that the popularity of the feature types changes slightly after this year in contrast to earlier years. Additionally, because we only have the first two quarters of 20222, the number of each kind is substantially lower than the others.

An outlier analysis was not performed because the company wanted to know the impact that COVID-19 had in the market of the footwear zippers.

4.4 Modelling

Now, we moved on to the modeling and prediction part of the study after observing certain peculiarities of our data. To achieve the project's goal, we used the last quarter of our data, dubbed *'2n trimestre 2022',* as the target variable and the other columns as the input variables. Moreover, for applying the models, the training and testing data was done with a split size of 0.8 for the train and 0.2 for the test data.

4.4.1 Binary Classification

The first apporach to do the modelling was to apply a binary classification whether the zippers were sold or not. In order to do the binary classification, we have overwrite the target variable with 0s and 1s. The 0s represents that the zipper was not sold according to our target variale, and the 1s represents that the zipper was sold. So, after this was applied, the balance of our target variable was:

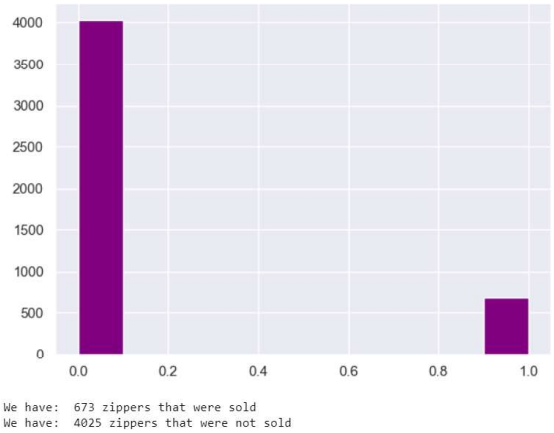


FIGURE 17: DISTRIBUTION BINARY CLASSIFICATION

In this graphic, we can observe that the binary classification is not really balanced, since we have a huge quantity of zippers that were not sold, and a low quantity of zippers that was sold. For this reason, a parameter in the function of the ML models was added. This parameter is called scale_pos_weight, and it works giving more imporatance to the class not dominant. In this case, the ratio given to the class number 1 is: $ratio = \frac{number\ of\ 0s}{number\ of\ 1s}$. The result was 5.98. This means that the models will give this quantity of importance in the not dominant class.

In the next table, we can see the results obtained from the different ML[2]:

	XGBoost	Light GBM	SVM	Gradient Boosting	Random Forest Classifier	CatBoost Classifier
Recall	0.95	0.93	1	0.98	0.97	1
Recall class 0	0.95	0.93	1	0.99	0.97	1
Recall class 1	0.71	0.74	0.08	0.35	0.59	0
Precision	0.95	0.96	0.87	0.90	0.93	0.85
Precision class 0	0.95	0.96	0.87	0.90	0.93	0.85
Precision class 1	0.69	0.62	0.85	0.71	0.79	0
Accuracy	0.91	0.90	0.87	0.89	0.92	0.85
F1 score	3.79	3.70	3.99	3.91	3.90	4
F1-score class 0	0.95	0.94	0.93	0.94	0.95	0.92
F1-score class 1	0.70	0.68	0.15	0.47	0.68	0
AUC score	0.83	0.83	0.54	0.66	0.78	0.5

TABLE 1: RESULTS BINARY CLASSIFICATION

Starting with Recall, all models perform well, with an overall Recall rate of at least 0.93. All models do reasonably well in class 0, with recall rates ranging from 0.93 to 1. However, the Recall rate for class 1 is much lower for all models, ranging from 0 to 0.74. We can see that SVM and CatBoost Classifier, are the models that have a highest recall for class 0 but almost none recall for class 1. This means that the models perfectly classify the zippers that will not be sold, but they do not have a good performance classifier the zippers that will be sold.

Moving on to Precision, all the models perform well, with a total precision ranging from 0.85 to 0.96. We can highlight that the CatBoost Classifier has the lowest precision rate in class 1 with a value of 0. Furthermore, we can also detect, that in general the precision for class 0 is higher

[2] Go to Appendix 10.2 to see the explanation of the metrics.

than the precision of class 1 in all the cases except for the SVM. This means that almost all the classifiers have a higher accuracy in identifying the positive instances for class 0 than for class 1.

Looking for the accuracy, we can see that the algorithms have a range from 0.85 to 0.92. The best model that achieves the highest accuracy is the Random Forest Classifier followed by the XGBOOST, while the lowest accuracy is achieved by the CatBoost Classifier.

For the F-1 score we can see that the values go between 3.70 and 4. The best F1-Score is achieved by CatBoost Classifier and the worst by the Light GBM algorithm.

Finally, we can see the AUC scores, of every algorithm. We can see that the best AUC score is obtained using the XGBOOST and the Light GBM and the worst case by the CatBoost Classifier, that is completely random.

According to the table, we can see that the top models that fits to this classification are: XGBOOST, Light GBM, Random Forest Classifier and Gradient Boosting. We can see from the table that the Random Forest Classifier and the Gradient Boosing have a good score for the total precision and for the individual precisions, but they have a lower individual recall for the class 1 compared with the Light GBM and the XGBOOST. Moreover, we can also detect that the AUC score for these two algorithms it is also lower than the other two models. So, in this case, the XGBOOST and the Light GBM are the models that performs better according to our goal. However, choosing between them it is hard since all the metrics obtained are close. Even so, as the metrics for the XGBOOST are generally higher than the Light GBM, I will consider that the XGBOOST give to us the best results for the binary classification of the zippers.

After doing the binary classification, as the main goal of this project is to predict the quantity of zippers that will be sold in the next sessions, the next approach was to apply regression models to the dataset.

4.4.2 Regression

For applying the regression models, the target variable was not anymore made of 0s (not sold) and 1s (sold). In this case, the exact quantity of sales of our target variable were used.

Before seeing the regression findings, a graph of the target variable's distribution was created. The zipper index is shown on the x-axis, while the number of zippers sold in the second quarter of 2022 is displayed on the y-axis. The following chart clarifies that there is a significant variation in the amount of zippers sold during the target variable. The plot can be seen here:

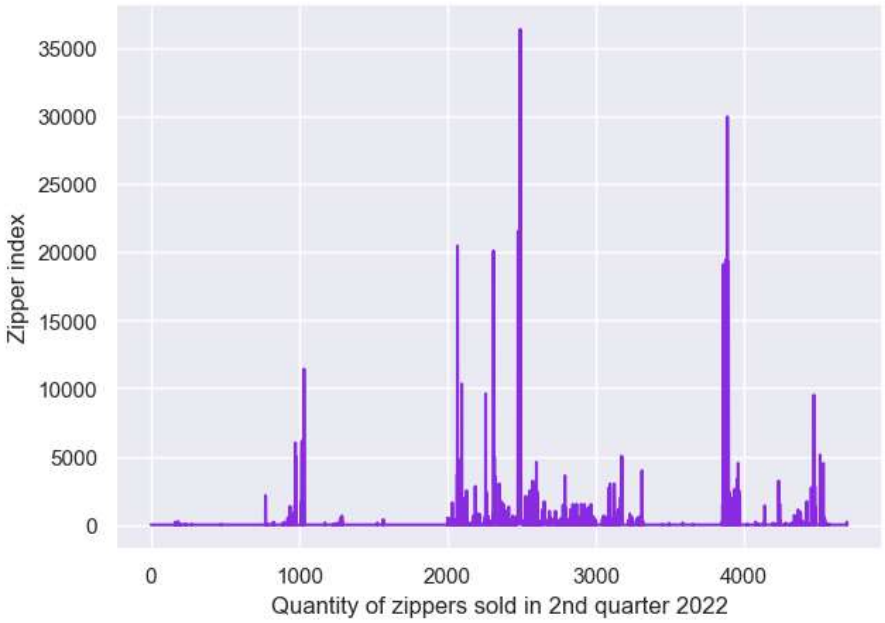


FIGURE 18: DISTRIBUTION OF THE TARGET VARIABLE

The first apporach using the regressors was to apply the different regressors using all the sales of the dataset. The results of this approach can be find in the next table[3]:

	Linear Regression	**CatBoost Regressor**
MSE	469001.72	1865370.69
MAE	236.93	284.43
RMSE	684.84	1365.79
R-SCORE	0.50	0.35
Standard Derivation of the relative error	Nan	Nan

TABLE 2: RESULTS REGRESSION WITH ALL THE DATA

From this table, we observe that the Linear Regression has a greater r-score than the CatBoost. This suggests that the Linear Regression fits our data better, although it is working with encoded data. Furthermore, we can observe how the MSE, MAE, and RMSE for both models are generally high, indicating that we made a significant error in projecting zipper sales. Also, it appears that the Linear Regression has lower values for these scores, implying that the regression error will

[3] Appendix 10.3

be smaller than the CatBoosr Regressor. However, we must keep in mind that the numbers are still large and far from the idyllic ones. Finally, we can observe that the standard derivation is Nans in both cases. This is because we are using all the dataset, including the zippers that were not sold.

Moreover, the scatter plots for this regression models were also plotted:

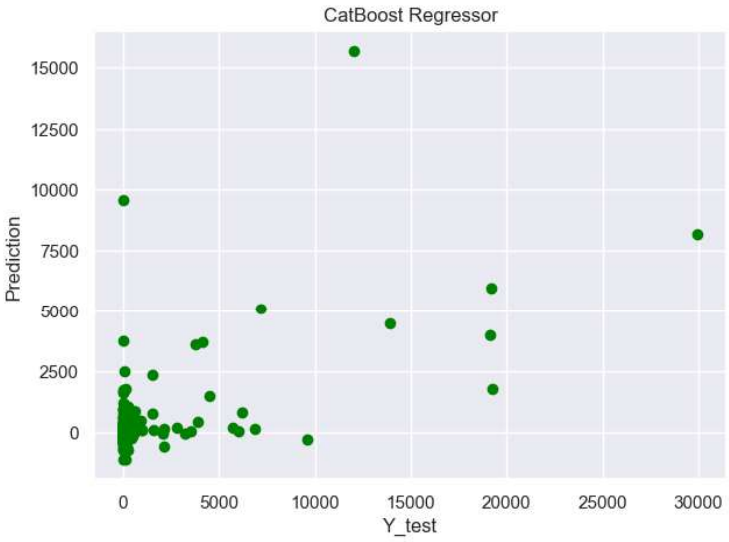


FIGURE 20: SCATTER PLOT OF CATBOOST WITH ALL DATA

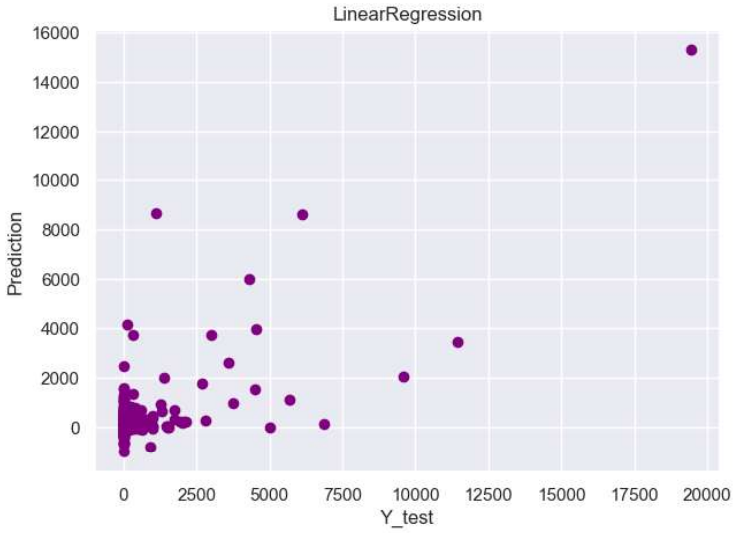


FIGURE 19: SCATTER PLOT OF LINEAR REGRESSION WITH ALL DATA

The x-axis of the scatter plot shows the real values from our test data, and the y-axis displays the prediction using the regression models. We can see that they are far away to construct a 45-degree line between the real value and the prediction, as the wished scatter plot. However, we can see that there is some correlation between the real value and our prediction since the r-score was not null. We can also highlight from the plots, that there is a relevant difference between the points of the plot. Most of the dots are around the zero and some of them have a rate between the prediction value and the real value far away from the zero. This is because we have a really different quantity of zippers sales in our target variable as we have already discussed in the beginning of this section. For this reason, another approach using regression models was done in order to find better results.

This other approach was done only taking into account the zippers that were sold according to our target variable. So, in this case, the regressor models were applied with a dataset of 673

rows and 25 columns, only taking into account the zippers sales grater than. The results of this new analysis can be find in the next table[4] :

	Linear Regression	**CatBoost Regressor**
MSE	2514403.46	12449618.64
MAE	863.10	1456.47
RMSE	1585.69	3528.40
R-SCORE	0.70	0.57
Standard Derivation of the relative error	104.92	38.68

TABLE 3: RESULTS REGRESSION WITH DATASET REDUCED

From the table, we can detect that we have obtained a good r-score from the Linear Regression. This means that 70% of the variance in the dependent variable can be explained by the independent variables. However, as we have a huge RMSE, this indicates that the average difference between the predicted values and the actual variables is still large. In other words, the results suggest that the Linear Regression has some predictive power but it may not be very accurate, because it may still be making large errors in its predictions. On the other hand, if we look in the CatBoost results, we can also see that the r-score has increased, but the RMSE, MAE and MSE have grown far more. Finally, we can also detect that the standard derivation is not null. This is due to the fact that we only had to execute the results with zippers purchased in the target variable, therefore no division by zero was performed. However, the standard derivatiom for both cases is still high and far away from our goal.

To sum up, we have seen that the Linear Regression fits better in our data but with higher errors. In order to get more accurate results, another approach was done.

4.4.3 Multiclass Classification

In this new approach, a multiclass classifier was used to try to get better results. The multiclass classification was only performed with the zippers that were sold in the last quarter of 2022. So, the modelling is done with 673 rows and 25 columns.

[4] Appendix 10.4 to see the Scatter plots.

The multiclass was done using four classes that are representing the quartiles of the target variable, which are:

- **1st quartile:** 100 zippers sold
- **2nd quartile:** 250 zippers sold
- **3rd quartile:** 752 zippers sold
- **4th quartile:** 36352 zippers sold

The distribution of the mulitclass is the following:

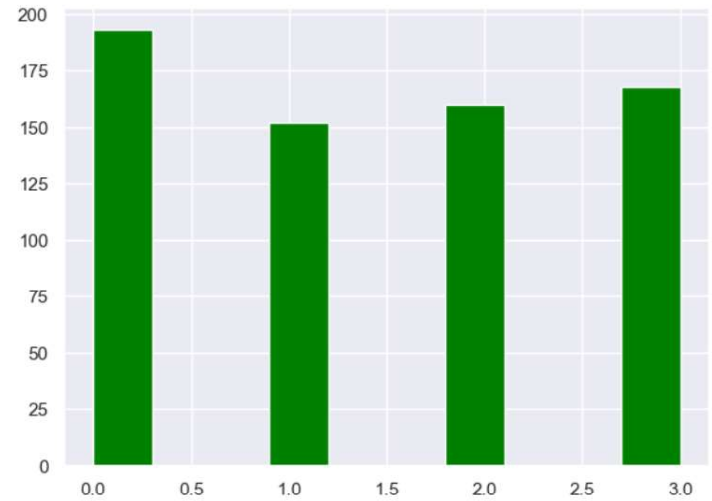


FIGURE 21: MULTICLASS DISTRIBUTION

We can highlight from the graphic that the fourth classes are quite balanced. However, we can see that the most zippers of the target variable were sold in the first class, which represents the first quartile. Since its not an unbalanced distribution the parameter to try to balance the classes was not added in the ML algorithms.

In the following table, we can find the results after applying the ML models:

	XGBOOST	Random Forest Classifier	SVM	Light GBM
Total recall	0.73	0.59	1	0.69
Recall 1st quartile	0.63	0.38	1	0.51
Recall 2nd quartile	0.33	0.24	0	0.28
Recall 3rd quartile	0.50	0.37	0	0.42

Recall 4th quartile	0.56	0.55	0.46	0.55
Total precision	0.75	0.58	0.56	0.6
Precision 1st quartile	0.59	0.44	0.32	0.49
Precision 2nd quartile	0.32	0.21	0	0.29
Precision 3rd quartile	0.45	0.34	0	0.41
Precision 4th quartile	0.71	0.55	0.77	0.58
Accuracy	0.52	0.38	0.36	0.45
Total F1 score	2.9	2.34	4	2.77
F1-score 1st quartile	0.61	0.40	0.48	0.5
F1-score 2nd quartile	0.33	0.23	0	0.28
F1-score 3rd quartile	0.47	0.35	0	0.42
F1-score 4th quartile	0.63	0.55	0.43	0.57

TABLE 4: RESULTS MULTICLASS

Starting with the analysis of the recall, we can see that values goes from 0.59 untill 1. The highest total recall it is achived for the SVM. However, if we look in the recall of the individual quartiles we can see that for the seoncd and third quartile, the recall is null. This means that the models perfectly classify the zippers that will be sold in the first quartile and also sometimes it will classify the zippers sold in the fourth quartile, but they will not classify to the other classes. Moreover, another highlight of this model that we can identify from the table, is that it has a really low total precision compared with the other models and also the precision for the classes two and three are zero. So, although these model has the highest total f1-score, we can clearly say that is not the ML model that we are looking for, since it is not identifying the zippers sold in the second and third quartile.

Furthermore, when we look at the precision findings without taking the SVM into account, we can see that the total precision increases from 0.58 to 0.75. Looking at the table, we can see that the XGBOOST, is the model that have the highest total precision and also the most balanced individual precision compared with the Light GBM and the CatBoost Classifier. From these two models, we can see that the precision for the class two and three that are much lower than the case of the XGBOOST algorithm. Moreover, from the XGBOOST, we can see that has the second one highest f1-score and the highest acuraccy score.

Looking for the results of the Random Forest Classifier, we can see that has the second best total recall and precision, and the individual recall and precision for the classes are better than the Light GBM and the SVM. However, if we comper the Random Forest Classifier with the XGBOOST, we can see that the XGBOOST performs better. To sum up, as we are interested to get the same importance to every class in order to find the best multiclass classification, we can see that the XGBoost is the model that is performing better in this multiclass classification.

4.5 Explainable and interpretability of ML

In this section, the SHAP plots based on the SHAP values[5] can be seen, in order to evaluate how the features of our data influences the ML models from the sections 4.3.1, 4.3.2 and 4.3.3.

Shap plots for the Binary Classification

As we already explained, the XGBOOST is the classification ML model that performs better for the binary classification. Here, we can find the plots to see which features have more importance during the application of the algorithm[6] :

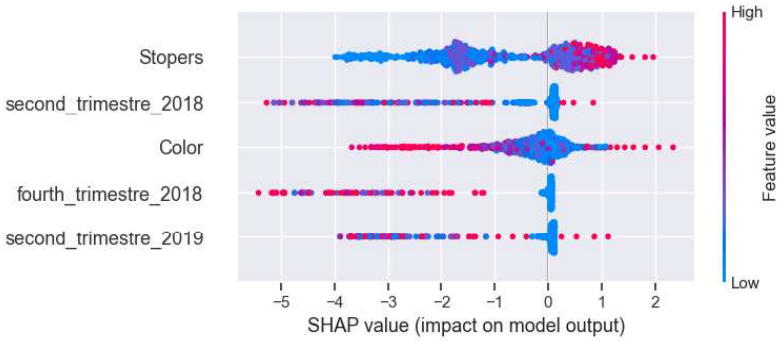


FIGURE 22: SUMMARY PLOT XGBOOST

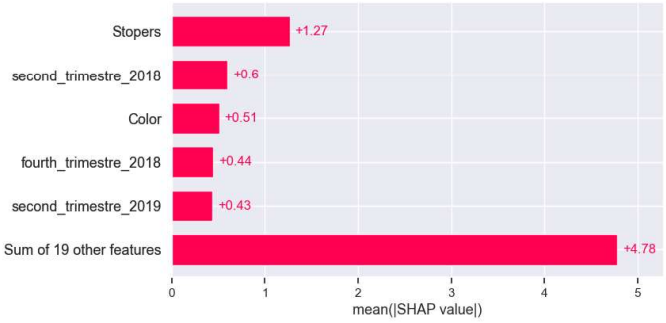


FIGURE 23: BAR PLOT XGBOOST

From this SHAP plots, we can see the first five attributes that are more important in determining whether the zipper will be sold in the binary classification. We may emphasize that the 'Stopers' are the most important characteristic for this model. Furthermore, stoppers labelled with low values will have a negative influence on the sale of the zipper, whilst stopers labelled with high values will have a favourable impact on the sale of the zipper. This means that the 'Stopers' labelled with low values increase the probability to be in the class number 0 (not sold), while the 'Stopers' labelled with high values increase the probability to be in the class 1 (sold). Moreover, for the feature 'Color', we can see that is the third attribute more relevant. We can identify that the high values will increase the probability of the zippers being part of the class 0 (not sold) while, and the low values labelled have will increase the probability for the zippers being in the class 1 (sold). If we look in the other features, it is harder to say how the low and the high values will impact in our model since in the plot they are more mixed. However, we can identify that for the 'second_trimestre_2018', 'second_trimestre_2019' and 'fourth_trimestre_2021', the

[5] Appendix 10.5 to see the explanation of the SHAP values and SHAP plots.
[6] Appendix 10.6 to see the complete SHAP plots for XGBOOST.

high number zippers will increase the probability to classify the zipper in the class 0, while the low quantity of zippers sold will increase the probability of the zippers being in the class 1.

Shap plots for Regression

The best model using the regression was Linear Regression with the reduced dataset. So, the explain ability of this model we can find it here[7]:

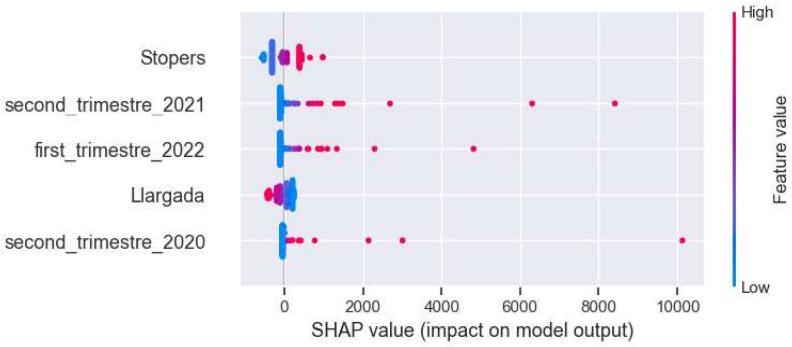


FIGURE 25: SUMMARY PLOT LINEAR REGRESSOR

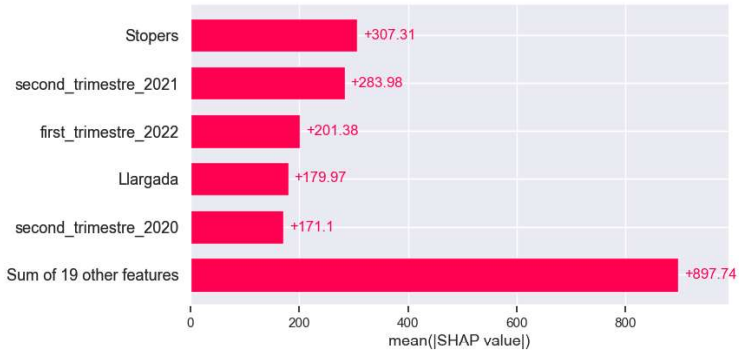


FIGURE 24: BAR PLOT LINEAR REGRESSION

The feature named 'Stopers' is the characteristic more relevant of the Linear Regressor model. In addition, we can also see the other fourth more relevant attributes from this model are: 'second_trimestre_2021', 'first_trimestre_2022', 'Llargada' and 'second_trimestre_2020'. First of all, we can identify that for 'second_trimestre_2021', 'first_trimestre_2022' and 'second_trimestre_2020', low zipper sales will increase the likelihood of the zippers being less sold, while high sales will produce a favourable impact to the probability of the zippers being more sold. Furthermore, we can observe that for the other two attributes, 'Stoppers' and 'Llargada,' the low labelled values enhance the possibility of the number of zipper being sold, while the high labelled values lower this probability.

Shap plots for the Multiclass

In this case, the shap plots for the Multiclass using the XGBOOST is the next one[8]:

[7] Appendix 10.7 to see the complete SHAP plots for the Linear Regression.
[8] Appendix 10.8 to see the complete plot.

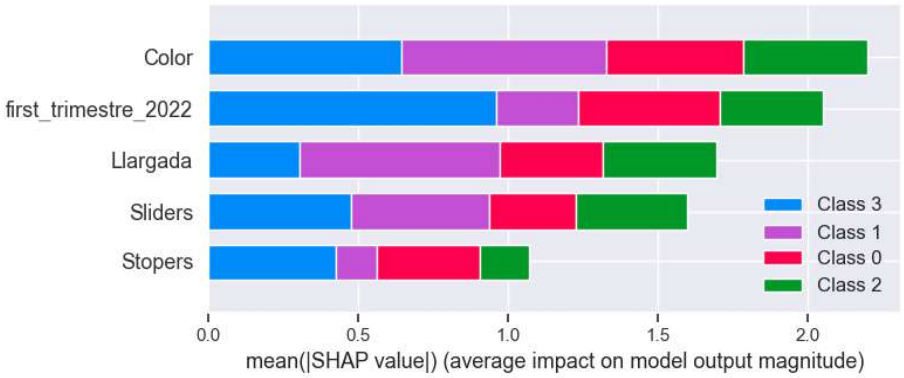


FIGURE 26: BAR PLOT MULTICLASS XGBOOST

The final figure shows the most relevant features for the multiclass categorization of zipper sales using the XGBOOST algorithm. In this case the five more relevant futures are: 'Color', 'first_trimestre_2022', 'Llargada', 'Sliders' and 'Stopers'.

In the preceding example, we can see that class 0 (first quartile) uses the characteristics 'Color,' 'first_trimestre_2022,' 'Llargada,' and 'Stoppers' more than the feature 'Sliders'. This means that for the class 0, the values of the most often utilized characteristics will enhance the likelihood of being correctly categorized. However, the variable 'Slider' reduces the chance of being correctly classified in this class.

Furthermore, we can see that the characteristics 'first_trimestre_2022' and 'Stoppers' are rarely utilized in class 1 (second quartile), whereas 'Color,' 'Largada,' and 'Sliders' are more commonly used. This implies that the chance of classifying correctly in the first class based on the variable's values of the 'first_trimestre_2022' and 'Stoppers' will drop. However, for the values of the features: 'Color', 'Llargada', and 'Sliders', being corrected classified in the second quartile will increase.

In the case of class 2 (third quartile), we can see that all the attributes will be used to classify the target variable.

Finally, the variable 'Llargada' is utilized less frequently in class 3 (fourth quartile) compared with the other features. In this situation, the values for this attribute will reduce the probability of obtaining a good classification for class 3.

5 STREAMLIT INTERFACE

In order to show the results to the company, an interface was build using the Streamlit library. In this link you can have acces to the web app: *https://judalmix-streamlit9-chiasa-webappmain-mdk10c.streamlit.app/.Zipper_prediction*

Requirements

The interface is crerated by three different pages, where in each page different actions can be done. The interface can support all the datasets with this format:

- The first ten columns of the dataset have to be object type with these exactly names: *'Codi', 'Descripció', 'Familia', 'Stopers', 'Sliders', 'Teeth', 'Color' 'Label', 'Llargada', 'Codi sense etiqueta'*.
- It must have a column named 'Total' with type integer that corresponds to the total amount of each zipper sold since the first month.
- The rest of the columns of the dataset, which correspond to the zipper sales for each month must be type integer.

Another requirement is that we must follow the order of the pages. Starting by the main, then to the Data Distribution page and finally to the Zipper prediction page. If the user does not follow this outline, the webapp will not work.

Navigating in the interface

The interface is created by three pages, which the names are:

- Main

Is the first page that we can see when the web page is open. It has a brief introduction about the goal of the interface and also explains the steps that the user has to do for the good operation of the webapp.

After this, there is a section where the user can upload the data of its interest as long as the data fullfies the requirements explained before. If the user upload a dataset with the good format, it will be able to see the dataset imported.

- Data Distribution

In this page, the user will be able to select an integer between 1 and 12 (both included) to group the months of the year. The user will be able to group the months of the year according to the prediction it wants to do. For example, if the user would like to obtain the prediction for the

next three months, the user should introduce the number 3. Here we can find the screenshot of this process:

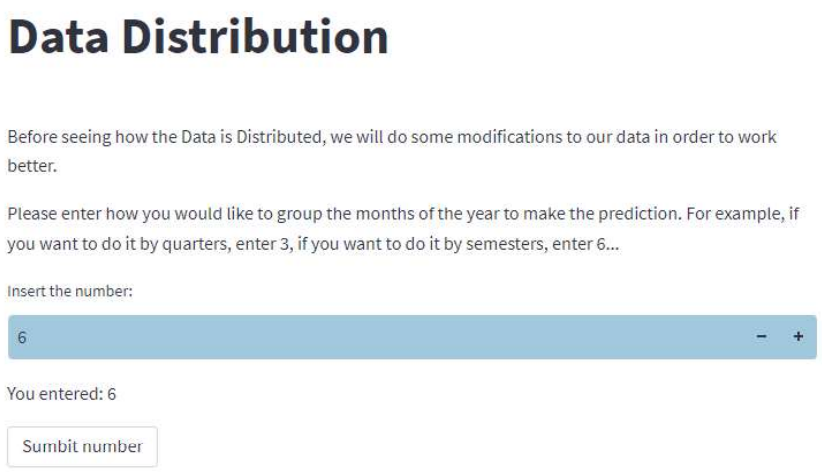
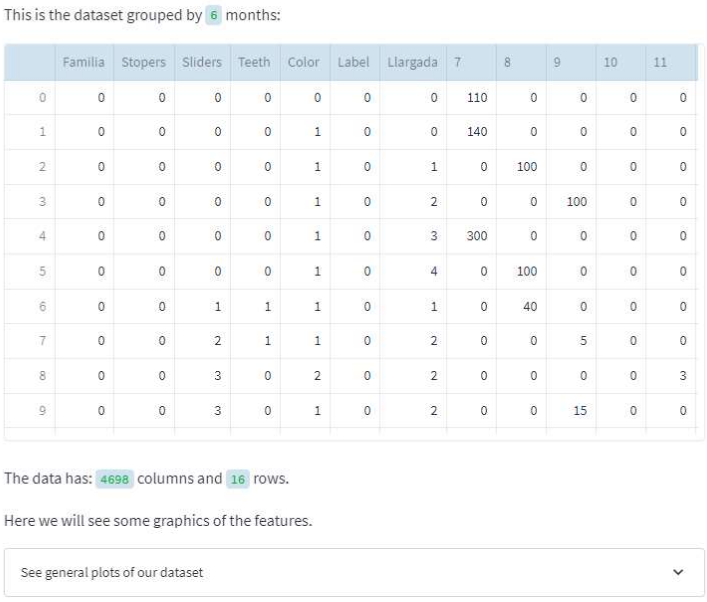


FIGURE 27: SCREENSHOT DATA DISTRIBUTION PAGE

In this case, the user chose to group the months of the dataset by semesters. So, after the user introduces the number, it must submit the number using the button of the interface called 'Sumbit number', and then the user will be able to see the dataset encoded with the months grouped. Here we can see for this specific example the dataset obtained:



	Familia	Stopers	Sliders	Teeth	Color	Label	Llargada	7	8	9	10	11
0	0	0	0	0	0	0	0	110	0	0	0	0
1	0	0	0	0	1	0	0	140	0	0	0	0
2	0	0	0	0	1	0	1	0	100	0	0	0
3	0	0	0	0	1	0	2	0	0	100	0	0
4	0	0	0	0	1	0	3	300	0	0	0	0
5	0	0	0	0	1	0	4	0	100	0	0	0
6	0	0	1	1	1	0	1	0	40	0	0	0
7	0	0	2	1	1	0	2	0	0	5	0	0
8	0	0	3	0	2	0	2	0	0	0	0	3
9	0	0	3	0	1	0	2	0	0	15	0	0

FIGURE 28: SCREENSHOT DATA MODIFIED EXAMPLE

As we can see, the data now is encoded, and the size of the dataset has changed since now columns are grouped by semesters.

Furthermore, this page also can show us some plots about the most popular features of our dataset. To see these plots the user only must go to the expander and select which plot it wants to see. Here we can see the result:

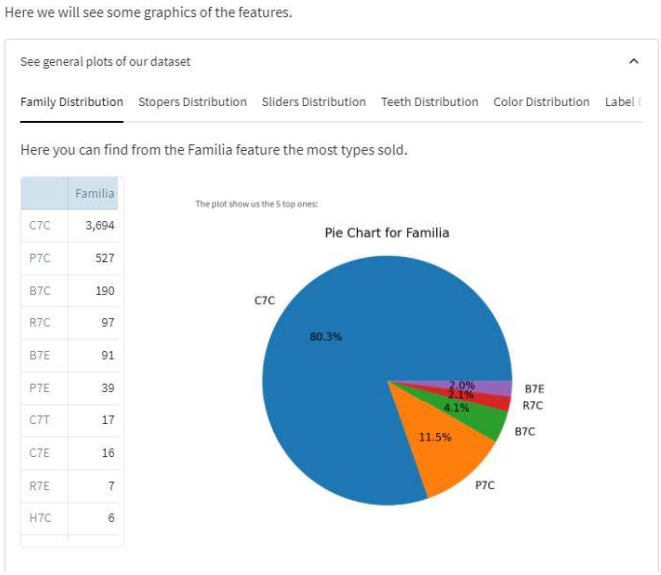


FIGURE 29: SCREENSHOT FEATURE DISTRIBUTION EXAMPLE

When the users arrive at this point, it must change to Zipper Prediction page.

- <u>Zipper Prediction</u>

In this page, the user can perform two different types of prediction and see the global interpretation and the individual SHAP plots for every model used in the prediction.
When the user goes to this page, first, he can find an explanation about what it can be perform in that space. After this the user can go to the two different expanders: Prediction expander and the explainable SHAP plots expander. Here you can find the screenshot of this:

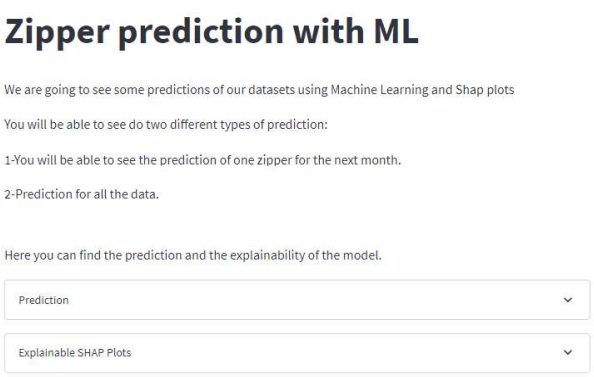


FIGURE 30: SCREENSHOT ZIPPER PREDICTION PAGE

If the user goes to the prediction expander, two different predictions can be performed:

The first one is to do a prediction for one zipper from the dataset. In this case, the user can choose the ID of the zipper that wants to predict and then the user can see the results of the prediction. For the predictions we have used the Multiclass classification and the Linear Regression with only the zippers that were sold according to our target variable.

As we already discussed in the section 4.4, the Linear Regression has a big error predicting, in this case if the error is huge, the results of the individual zipper prediction only will consider the MULTICLASS prediction. Here you can find an example of the process:

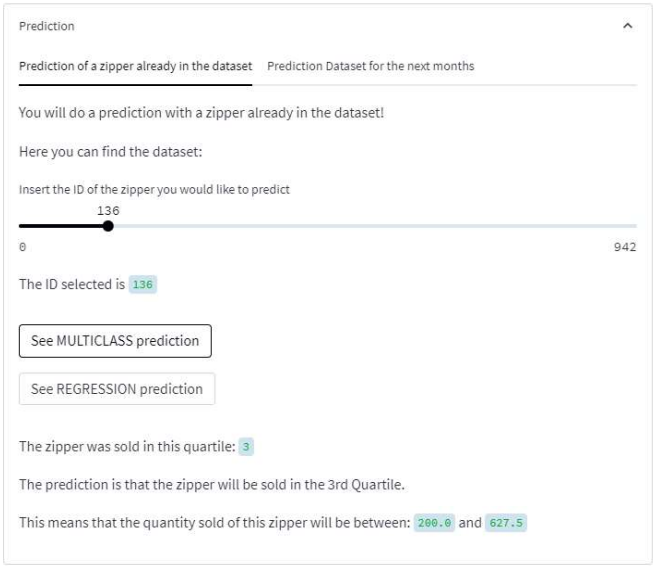


FIGURE 31: SCREENSHOT EXAMPLE MULTICLASS PREDICTION

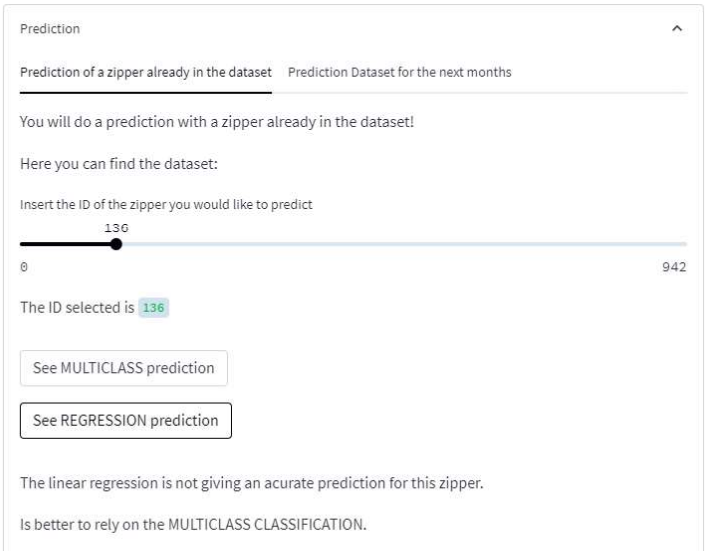


FIGURE 32: SCREENSHOT EXAMPLE REGRESSION PREDICTION

In the screenshot of the left, we can see that the user selected the ID number 136. In this case, we can see that the multiclass prediction says that it will be sold in the third quartile between 200-637 zippers for this specific case. On the other hand, the screenshot of the right we can see that the Linear Regression is not performing well since the prediction shows that we must take only into account the prediction of the multiclass.

The second prediction that can be done is to do the prediction for all the zippers of the dataset for the next semesters using the Linear Regression. In this case, the user must know that the result might have errors since the big errors that we have seen in the section 4.4.2.

On the other hand, if the user goes to the Explainable Shap plots expander, the user can choose to see the Shap plots of the multiclass classification or the Linear Regression Classification. In

the case of the Multiclass classification, the user will only see the global explication of the shap plots already discuss in the section 4.5 and for the Linear Regression, the user will be able to see the global and the individual waterfall plot. In the case the user chooses the Individual plot, it will have to introduce the ID of the test data that wants to see how the model is performing. Here we can find an example:

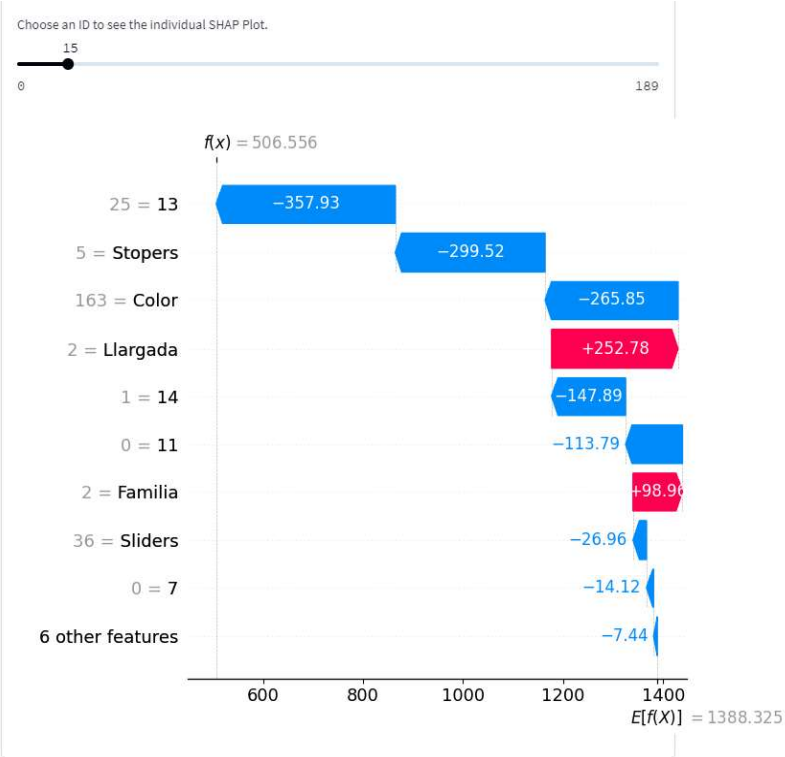


FIGURE 33: WATERFALL PLOT EXAMPLE LINEAR REGRESSION

In this instance, the user selected the ID of the test data number 15. As we can see, the x-axis shows shape values, and the y-axis represents the most significant feature for the ID chosen from the output model's test data. Next to the feature name, we can also see different values for each characteristic. The most critical variables for the ID picked that will have a negative influence are column 13, 'Stoppers,' 'Color,' column 14, and column 11. This indicates that these traits, along with their values, will reduce the likelihood of the zipper being sold. However, we can see that the variables 'Llargada' and 'Familia' will have a favourable effect. This means that adopting these features with their values increases the likelihood of the zipper being sold.

After all this process, the user will obtain the prediction and it will be able to see some explain ability of how the ML models work through the different features of the data imported.

6 CONCLUSIONS

Evaluating the impact that the differents ML algorithms have in the data it is essencial for every development of Machine Learning project. The goal of the project was to find a good model for the sales forecast to help Kreband company to reduce costs and improve efficiency. To achieve the main goal, the MLOPS process was fundamental in order to know the outline and the next steps.

Specifically, using different visualitzation techniques in the beginning of the project, had helped us to see how the footwear zippers sales was distributed and get an outline of the data and about which type of features were more sold in the past years.

Moreover, developing different ML approaches with different types of algorithms and to analyze the results were also performed correclty. The first approach was to apply a binary classification in order to see how the algorithms worked classifing between the zipper solds or not. Secondly, different types of regression were applied to see how the regressors worked trying to predict the amount of zippers sold. In this section, we have seen that the Linear Regression was the best model fitting our data, but the error of the predictive results was high. Finally, the last approach was done using a multilcass classification by the quartils of the zippers sold. In this case, we have seen that the XGBOOST was the best model and it classifies more then the fifty percent correctly. On the other hand, all these techniques were evaluate using their respectively metrics and also the weak and strongs points of each algorithm were developed.

Furthemore, seeing the importance that the features have in the best ML models of every approach was also achived using the shap values and the shap plots. Using this new plots, we were able to see the most relevance features and how they impact in the ML model.

Finally, an interface was developed in order to tave a visualitzation tool that can help us to understand the results. In this interface, the best ML models obtained were applied to see how performed with real cases.

7 FUTRUE WORK

After seeing the different results of our project, we can think in a further investigations for this project. Here we can find two different ideas for the next studies:

The first idea for a further analysis might be considering that our data it is stationary. In this case, a time series analysis can be applied to see a new approach to apply in our data and see more pattrens of the zipper sales.

On the other hand, the second idea for the further work, is applying Deep Learning (DL) techniques as Recurrent Neuronal Networks (RNN). Using this new techniques can be another approach to continue with the investigation of the zippers sales and how are related with the fashion trends.

8 BIBLIOGRAPHY

[1] *Who we are. (n.d.). Chiasa. Retrieved June 14, 2023, from https://www.chiasa.es/qui-som/*

[2] *Structure of a Zipper / YKK FASTENING PRODUCTS GROUP. (n.d.-b). YKK FASTENING PRODUCTS GROUP. https://www.ykkfastening.com/products/zipper/s_zipper.html*
[3] *Wikipedia contributors. (2023). Zipper. Wikipedia. https://en.wikipedia.org/wiki/Zipper*

[4] *What is XGBoost? (n.d.). NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost/*

[5] *Gong, D. (2022, July 12). Top 6 Machine Learning Algorithms for Classification. Medium. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501*

[6] *XGBoost Documentation — xgboost 1.7.5 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/*

[7] *Wikipedia contributors. (2023b). Random forest. Wikipedia. https://en.wikipedia.org/wiki/Random_forest*

[8] *IBM Documentation. (n.d.). https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-howsvm-works*

[9] *Hoare, J. (2022, August 24). What is Gradient Boosting? - Gradient Boosting Explained - Displayr. Displayr. https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/*

[10] *Wikipedia contributors. (2023a). LightGBM. Wikipedia. https://en.wikipedia.org/wiki/LightGBM*

[11] *Gupta, A. (2023). The Best Guide to Understand Regression Analysis in Excel. Simplilearn.com. https://www.simplilearn.com/tutorials/excel-tutorial/regression-analysis*

[12] *Wikipedia contributors. (2023b). Linear regression. Wikipedia. https://en.wikipedia.org/wiki/Linear_regression*

[13] *Wikipedia contributors. (2023b). Confusion matrix. Wikipedia. https://en.wikipedia.org/wiki/Confusion_matrix*

[14] Jayaswal, V. (2021, December 15). Performance Metrics: Confusion matrix, Precision, Recall, and F1 Score. Medium.
https://towardsdatascience.com/performance-metrics-confusion-matrix-precisionrecalland-f1-score-a8fe076a2262

[15] Classification: Accuracy. (n.d.). Google for Developers.
https://developers.google.com/machinelearning/crash-course/classification/accuracy

[16] Zach. (2021). How to Interpret Root Mean Square Error (RMSE). Statology.
https://www.statology.org/how-to-interpret-rmse/

[17] GeeksforGeeks. (2023). R squared in Regression Analysis in Machine Learning. GeeksforGeeks.
https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/

[18] Error absolut i realtiu. (n.d.).
http://formulesmatematiques.blogspot.com/2012/10/error-absolutrealtiu.html

9 APPENDIX

9.1 Appendix 1

ML models are data-trained algorithms that produce predictions or choices. These models are intended to detect patterns or relationships in data and then utilize that information to generate predictions or classify fresh data based on its properties. In this project, Supervised Learning (SL) Models and Deep Learning (DL) were employed to construct the project.

Supervised Learning Models

SL models are a sort of machine learning model that is trained on labelled datasets with known outcomes. SL models are trained using a training dataset, which includes the input data along with the corresponding output labels. Then, the model learns to identify patterns or relationships in the data that will enable it to make accurate predictions on new and unseen data [4]. After training, the model is assessed using a validation dataset to determine its accuracy and generalization performance. If the model performs well on the validation dataset, it can be used to forecast new, previously unseen data.

Classification and regression models and are the two types of SL models that were used to develop the project.

- **Classification Models**

Classification models are SL models that predict the category or class of a new input based on the characteristics or features of the input. During the training phase, classification algorithms learn from labelled data. The model is fed a collection of input data along with their labels and learns to find patterns or correlations between the data and their labels. Once trained, the model can be used to categorize new input data based on its attributes. [5]

To create the project, categorization algorithms were employed to determine whether the zippers will be sold. To make this prediction, we can utilize a variety of methods, which we can categorize based on the types of features they can handle:

- Extreme Gradient Boosting (XGBOOST)

XGBoost works by training decision tree models iteratively to correct the prior model's faults, with each subsequent model focused on the most difficult-to-classify samples. Using this method, XGBoost can obtain excellent accuracy even on complex datasets with high-dimensional feature spaces [6].

- Random Forest (RF) Classifier

The Random Forest classifier works by generating several decision trees from different subsets of the training data and characteristics. RF classifier makes predictions using an ensemble of decision trees rather than a single decision tree. The last prediction is made by merging all the different tree predictions. This approach reduces overfitting by lowering the likelihood of any decision tree overfitting to the training data [7].

- Support Vector Machines (SVM)

SVM works by mapping data to a high-dimensional feature space and then constructing a hyperplane to split distinct classes, in this case, predicting whether zippers will be sold [8]. The goal of this method is to apply support vectors to find the best hyperplane with the greatest margin between classes. This finds the points closest to the hyperplane, which defines which class our point belongs to.

- Gradient Boosting

Gradient boosting is a ML approach that combines weak models, typically decision trees, to generate a strong model [9]. It trains these weak models in stages, with each model correcting the mistakes of the previous model. This procedure is repeated until the model achieves maximum accuracy.

- Light GBM

LightGBM is a powerful and efficient gradient boosting framework based on decision tree methods for machine learning [10]. When training a model with this algorithm, the framework starts with an initial decision tree and uses gradient boosting to build additional trees that correct the errors of the previous tree. LightGBM uses a leaf-wise approach to tree growth, which means that the tree is split based on the maximum gain in a single leaf, rather than splitting all the nodes of a level at once. This approach creates taller and thinner trees, which helps to reduce overfitting and improve performance.

- CatBoost Classifier

CatBoost Classifier trains an ensemble of decision trees using gradient boosting. It can handle both numerical and categorical characteristics and employs many anti-overfitting strategies, including as random data permutations and bagging. During training, CatBoost modifies the weights of features based on their value to the model, giving greater weights to more informative characteristics. CatBoost employs the ensemble of decision trees to generate predictions on new data after the model has been trained.

- **Regression models**

Regression models are statistical approaches for estimating the connection between a dependent variable and one or more independent variables [11]. The goal of regression analysis

is to find the best fit line or curve that shows the relationship between the variables and can be used to predict.

During the thesis's development, two different regression models were utilized, after the classification approach, solely with the zippers that were sold to do a higher precision trying to forecast the quantity of zippers that will be sold.

The regression models used are:

- Linear regression

The linear regression seeks the best-fit line that describes the relationship between the independent factors and the variable. It entails determining the best fit line from a set of data points that illustrates the linear relationship between the two variables. The best fit line is obtained by minimizing the sum of the squared vertical distances between each data point and the line [12]. In our scenario, the target variable will be the fourth quarter of 2022, and the other factors in our dataset will be the independent variables attempting to forecast the best link between them.

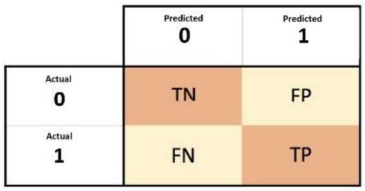
- CatBoost Regressor

CatBoost regressor is a ML technique that performs regressions tasks using gradient boosting decision trees without the need of pre-process the data. This means that can handle categorical and numerical data.

9.2 Appendix 2

To analyse how the ML models performs with the data, some metrics and plots were used:

- **Confussion matrix:** A confusion matrix is a table used to evaluate the accuracy of a classification model. It displays the number of correct and incorrect predictions made by the model for each class [13]. The matrix is composed by four values: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).



- **Precision:** The precision value all the positive predicted, what perentage is truly positive. It lies between 0 and 1, and we are interested that the model give to us a precision close to 1. It can be computed using this formula [14]: $Precision = \frac{TP}{TP+FP}$
- **Recall:** The recall it is out of the total positive, the percentage that are predicted positive. It lies between 0 and 1, and usually we are intersted in obtaining values around 1. It can be computed using [14]: $Recall = \frac{TP}{TP+FN}$

- **Acuraccy:** The accuraccy score is the fraction of predictions that the model got right [15]. The acuraccy lies between 0 and 1, and we are intersted that our model give to us values around 1. It is computed using the formula: $Acuraccy = \frac{TP+TN}{TP+TN+FP+FN}$
- **F1 score:** F1 score is the harmonic mean of the precision and recall. As bigger is the value obtained in the f1-score better is the performance of the mode. It can computed using the formula: $F1\ score = \frac{2*(precision*recall)}{(precision+recall)}$
- **Receiver Operating Characteristic (ROC) score:** The ROC curve is a graphical representation of a binary classification model's diagnostic capacity when its discrimination threshold is modified. It is calculated by graphing the true positive rate (TPR) versus the false positive rate (FPR) at different threshold values. A ROC socre close to 1 indicates better performing for the model, and a ROC socre close to 0 indicater a bad performance.

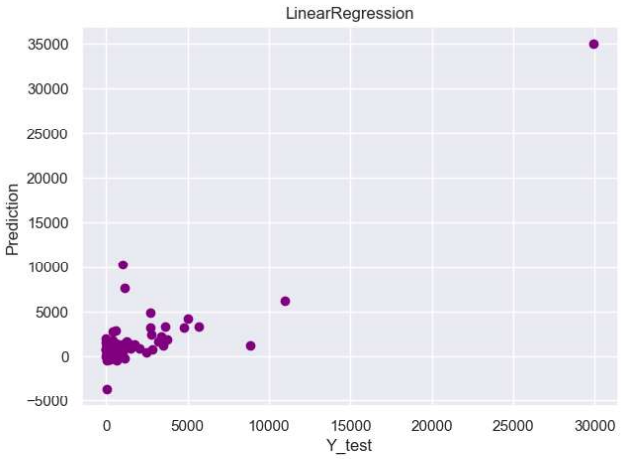
9.3 Appendix 3

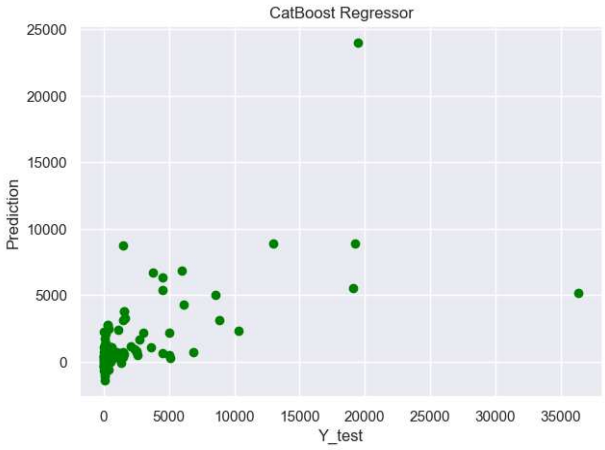
The metrics for the regression models are:

- **Root Mean Squared Error (RMSE):** It is a metric used in ML analysis to calculate the difference between the actual and predicted values. It represents the average distance between predicted and actual values and serves as a measure of the model's accuracy. Lower RMSE values imply that the model is better fitted to the data [16].
- **R-squared(R2):** R-squared is a statistical measure used to assess a regression model's goodness of fit. R-square ranges from 0 to 1, with 1 indicating that the model fits the data perfectly and 0 indicating that the model does not explain any variability in the target variable [17].
- **Scatter plot:** A scatter plot is a graph that displays numerical data points as dots on a two-dimensional Cartesian coordinate plane. In our case, the horizontal axis is our test data, and the y-axis the prediction of the regression. Using this information, we can visualize the relationship between the two variables, allowing us to better understand the data and detect any trends or patters that may exist.
- **Mean absolute error (MAE):** It is computed by averaging the absolute differences between predicted and actual values. MAE is frequently used in the evaluation of regression models and time series forecasts to assess prediction accuracy. The lower the MAE, the more accurate the model or forecast.

- **Relative error:** Relative error is a measure of the accuracy of an approximation or prediction compared to the true value. It can be computed using the formula: $Relative\ error = \frac{Absolute\ error}{Real\ value}$. Relative error provides a way of comparing the accuracy of predictions or estimates across different scales, as it corrects for differences in the magnitudes of the values being compared [18].

9.4 Appendix 4





9.5 Appendix 5

The shap approach is a strategy for explaining how each feature influences the model and for performing local and global analysis on the dataset. This is achieved using the SHAP values and the SHAP plots.

The SHAP values are a means to explain the output of machine learning models and comprehend the significance of various model features. SHAP values are used to credit the contribution of each feature to the final prediction by averaging the changes in model output when the feature is included or excluded in all feasible combinations. To compute SHAP values for a given machine learning model and data point, we first compute the contribution of each feature by comparing expected output when the feature is present vs anticipated output when the feature is absent. This is done for all possible feature combinations in the given dataset. The Shapley value, which is a measure of each feature's contribution to the overall forecast, is then calculated by averaging the contributions across all conceivable feature combinations. This technique has been designed to satisfy key axioms such as consistency, symmetry, and additivity. The generated SHAP values represent each feature's contribution to the model's output for the given data point.

On the other hand, we have the SHAP plots, that are graphs that depic the SHAP values. They descibre the output of ML models and they help users to understand how each attribute contributes to the overall forecast. In this project, the next SHAP plots were used:

- Waterfall plot

The figure depicts the positive and negative aspects that contribute to the forecast, with the most essential features at the top and the least important at the bottom.

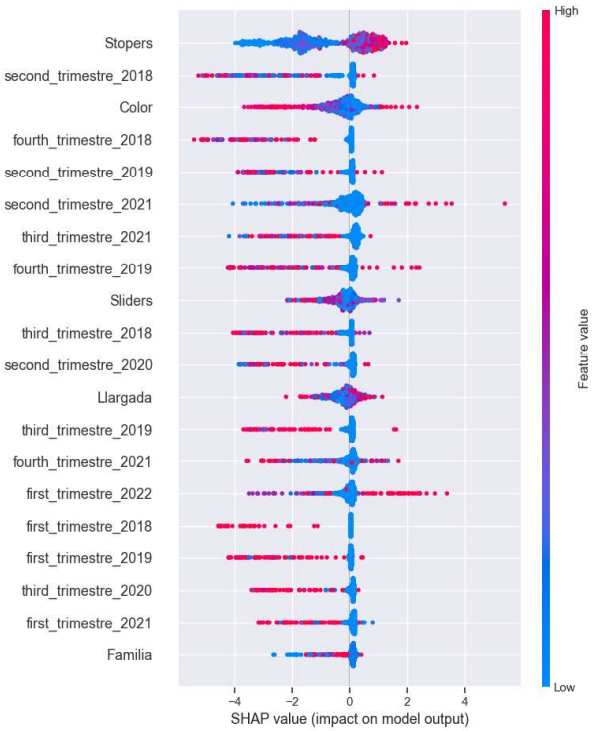
- Bar plot

The average or sum of the absolute SHAP values for each feature in a ML model is displayed. It depicts the relative relevance of each feature in the model, with the most important element at the top and the least important at the bottom of the plot.

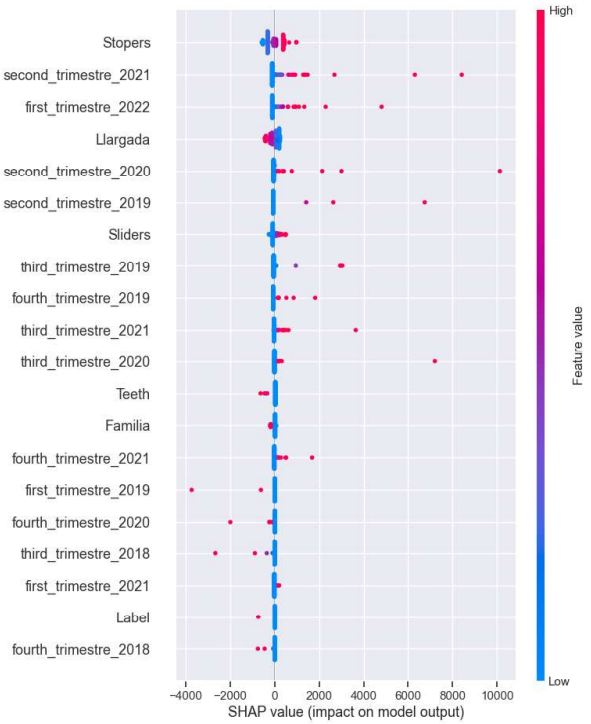
- Summary plot

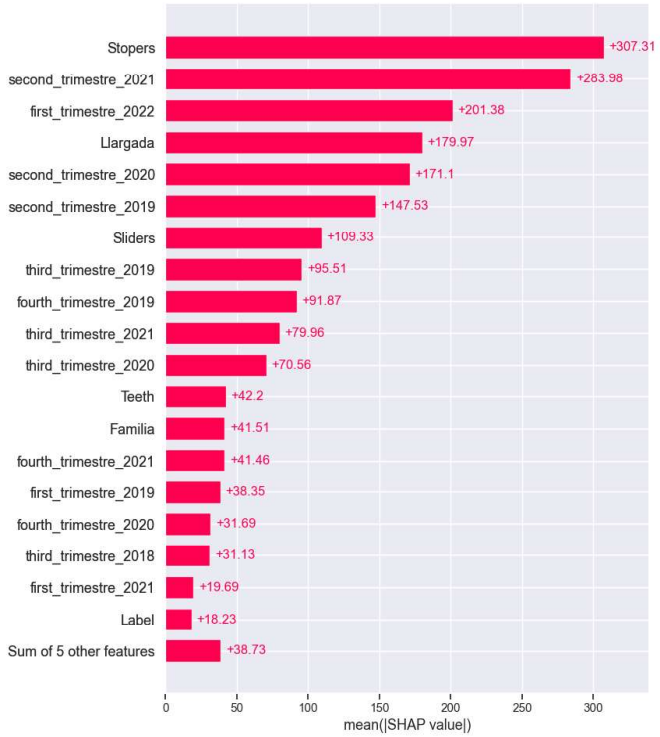
It is a graphical depiction of a ML model's global feature relevance. It demonstrates how much each feature contributes favorably or negatively to the model's output and can aid in the identification of patterns and linkages between features.

9.6 Appendix 6



9.7 Appendix 7





9.8 Appendix 8