Deloitte. Digital

ETSEIB UPC

Customer Churn Analysis In The Automotive Industry Utilizing Logistic Regression And CRISP-DM Methodology

24 MAY 2022

PEDRO CAROL ESCOLÁ

Context about the project

Project Origin

This project originated from work done at Deloitte Consulting, where support was provided for the marketing campaigns of a client in the automotive sector

Motivation

The motivation for this project comes from the client's lack of knowledge about their customers' behavior and my proactivity in learning new advanced analytics techniques

Objectives

The objective of this project is to identify customers who are at a higher risk of abandoning the brand by applying the CRISP-DM methodology

KEYNOTE STRUCTURE

DATA STUDY AND UNDERSTANDING

- Data Structure
- Data Extraction
- Data Preparation

DATA ANALYSIS AND FEATURE SELECTION

- Customer profiling
- Correlations between variables

MODELING

- Logistic Regression
- Evaluation methods
- Evaluation metrics
- Over-Sampling
- Model application and evaluation

CONCLUSIONS

- Objective evaluation
- Personal conclusion
- Future work

KEYNOTE STRUCTURE

DATA STUDY AND UNDERSTANDING

- Data Structure
- Data Extraction
- Data Preparation

DATA ANALYSIS AND FEATURE SELECTION

- Customer profiling
- Correlations between variables

MODELING

- Logistic Regression
- Evaluation methods
- Evaluation metrics
- Over-Sampling
- Model application and evaluation

CONCLUSIONS

- Objective evaluation
- Personal conclusion
- Future work

Data Structure

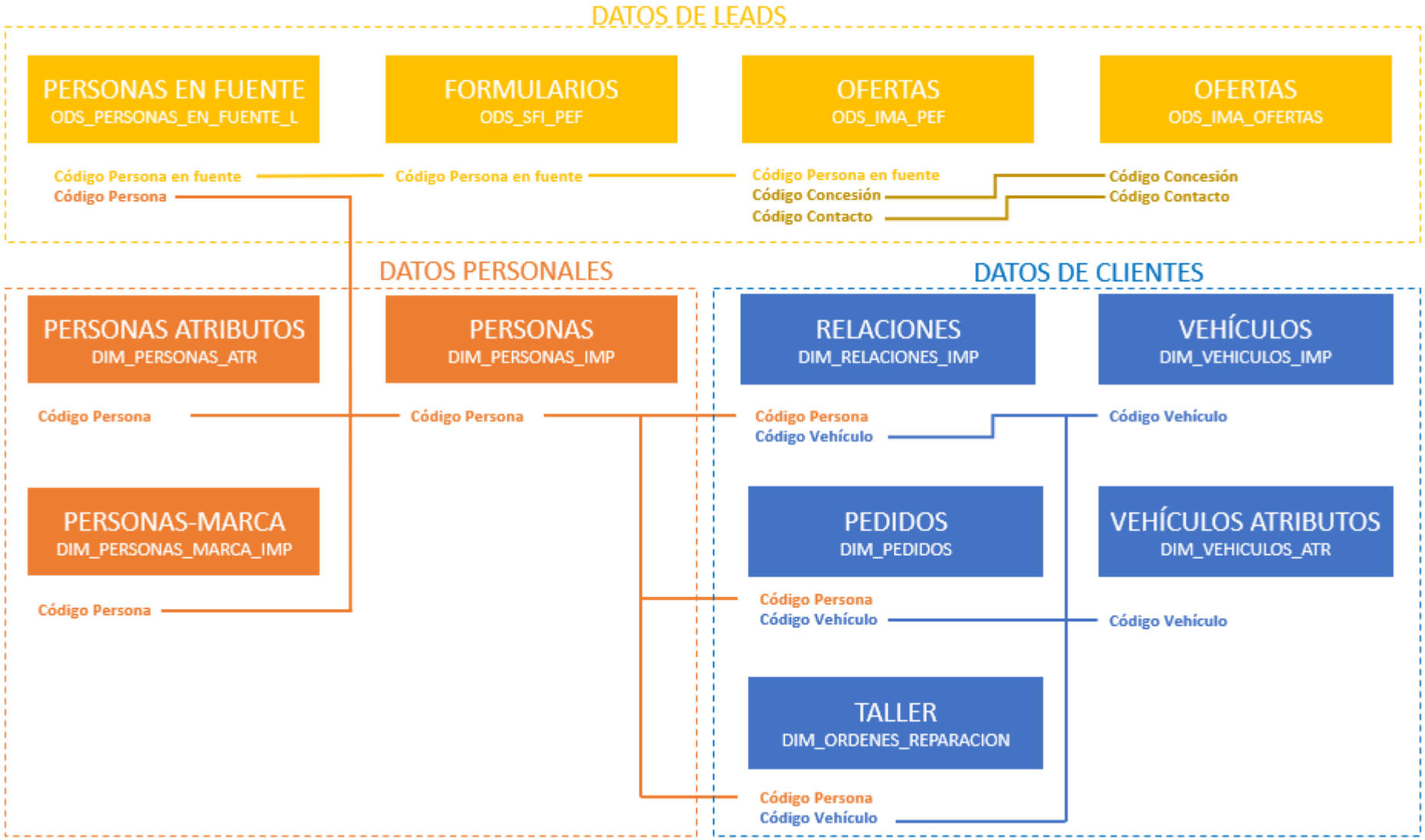


*Illustration names might be in Spanish since the project was originally documented in Spanish

Data Extraction

Data extraction involves 3 separate SQL queries on the tables for personal data, client data, and lead data.

Extraction of Clients

The extraction of clients consists of the personal data and client data tables, excluding the workshop visits table

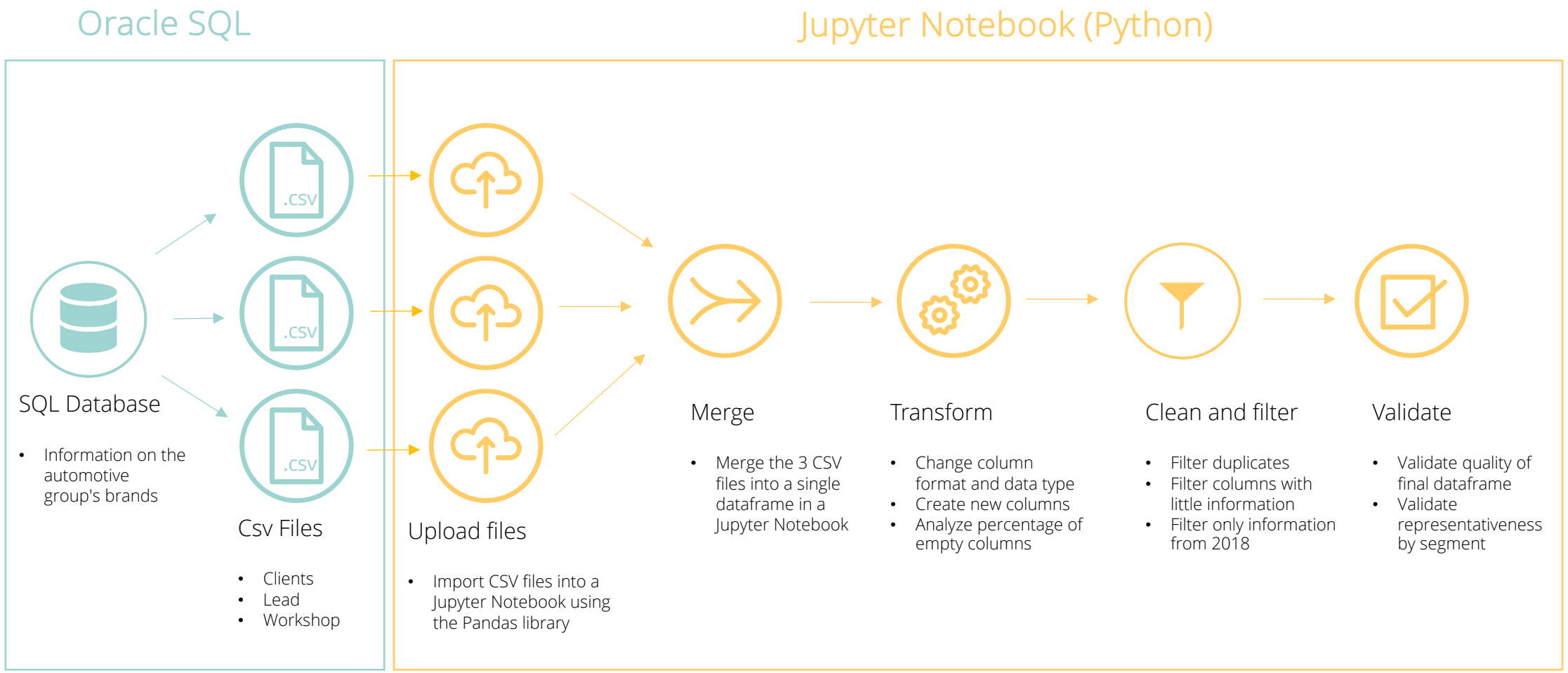
Extraction of Leads

The extraction of leads consists of the personal data and lead data tables. Lead data is extracted in an aggregated manner due to the high volume of data in transactional tables

Extraction of Workshop visits

The extraction of workshop visits consists of the personal data and client data tables. From the client data, only the workshop table is included. Data from the workshop table is extracted in an aggregated manner due to the high volume of data in transactional tables

Data Preparation



KEYNOTE STRUCTURE

DATA STUDY AND UNDERSTANDING

- Data Structure
- Data Extraction
- Data Preparation

DATA ANALYSIS AND FEATURE SELECTION

- Customer profiling
- Correlations between variables

MODELING

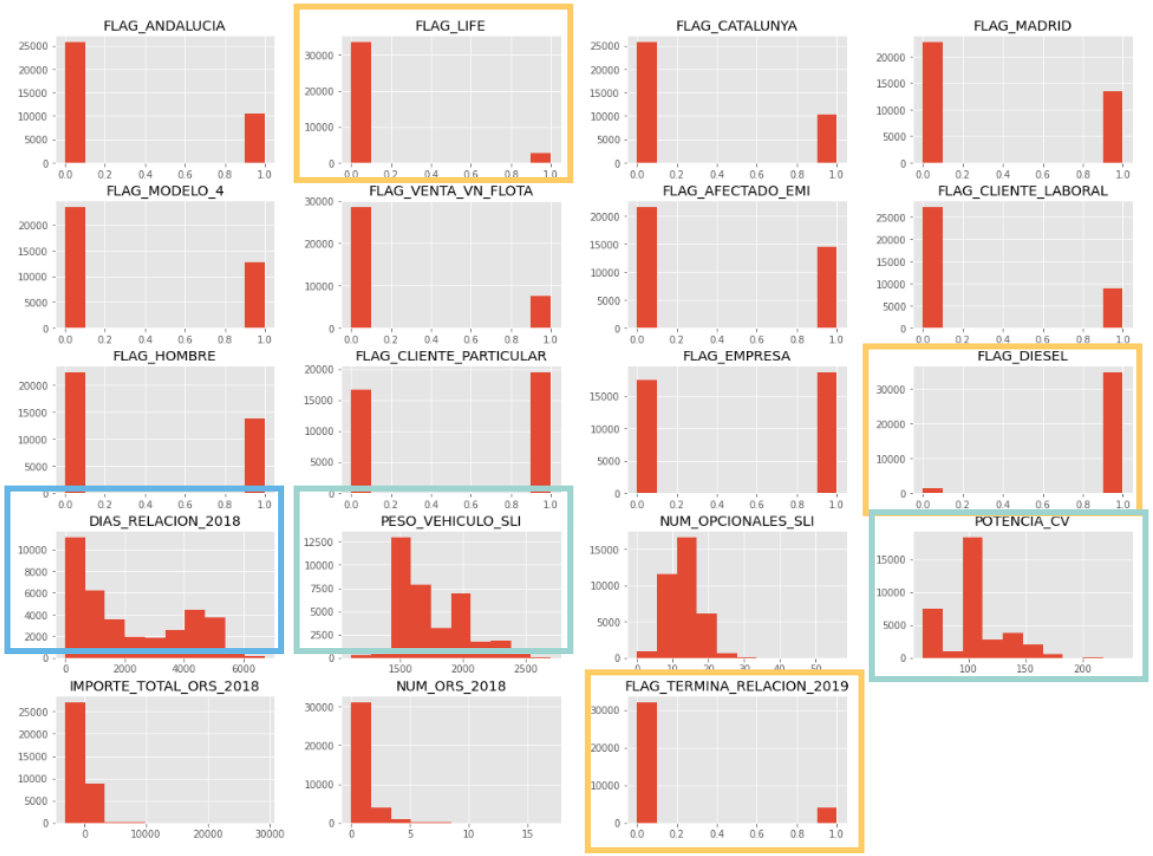
- Logistic Regression
- Evaluation methods
- Evaluation metrics
- Over-Sampling
- Model application and evaluation

CONCLUSIONS

- Objective evaluation
- Personal conclusion
- Future work

Customer Profiling

We analyze all the variables of our final dataframe with the customer data from the year 2018



- We observe that most of the variables are fairly balanced
- The variables FLAG_LIFE, FLAG_DIESEL y FLAG_TERMINA_RELACION_2019 are the ones that are most imbalanced
  - Workers with diesel vehicles who do not end their relationship in 2019 would be the majority profile in our dataset
- The variables POTENCIA_CV and PESO_VEHICULO show that there are two subgroups. Upon analysis, we have confirmed that these two subgroups correspond to the vehicle models studied (model 4 and model 5)
- The variable for days of relationship in 2018 also shows two subgroups, but these are not related to the vehicle model

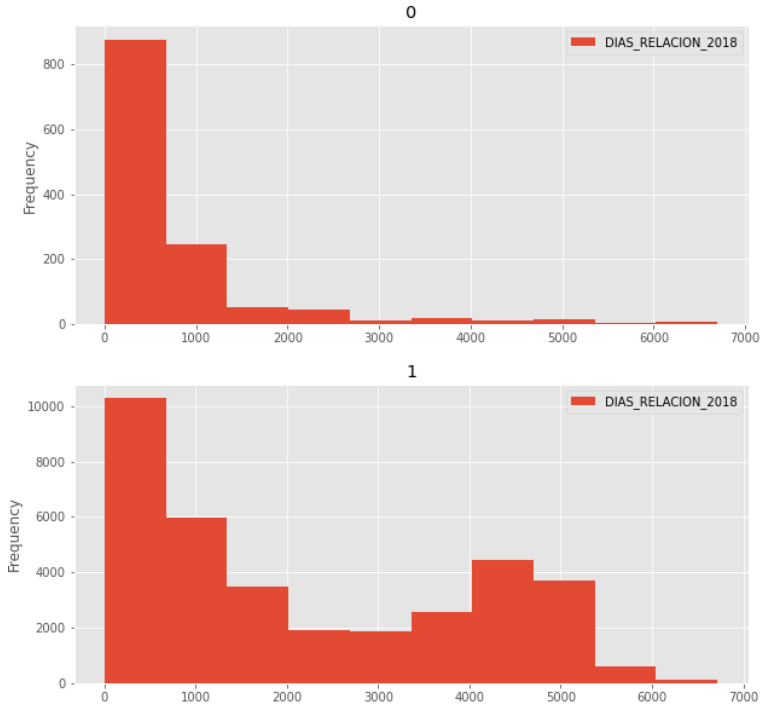
Correlations between variables

No significant correlations are observed for the predictor variables

Correlation with the variable to be predicted

DIAS_RELACION_2018	-0.098015
FLAG_HOMBRE	-0.068684
FLAG_ANDALUCIA	-0.063188
FLAG_CLIENTE_LABORAL	-0.032565
FLAG_DIESEL	-0.015934
FLAG_CATALUNYA	-0.011920
FLAG_CLIENTE_PARTICULAR	-0.004944
FLAG_MODELO_4	-0.000955
FLAG_VENTA_VN_FLOTA	-0.000853
POTENCIA_CV	0.016959
PESO_VEHICULO_SLI	0.021044
FLAG_LIFE	0.021318
NUM_OPCIONALES_SLI	0.028377
IMPORTE_TOTAL_ORS_2018	0.030184
NUM_ORS_2018	0.043881
FLAG_MADRID	0.074054
FLAG_AFECTADO_EMI	0.080502
FLAG_EMPRESA	0.084292
FLAG_TERMINA_RELACION_2019	1.000000

Diesel vehicle behaviour vs. non-diesel vehicle behaviour



KEYNOTE STRUCTURE

DATA STUDY AND UNDERSTANDING

- Data Structure
- Data Extraction
- Data Preparation

DATA ANALYSIS AND FEATURE SELECTION

- Customer profiling
- Correlations between variables

MODELING

- Logistic Regression
- Evaluation methods
- Evaluation metrics
- Over-Sampling
- Model application and evaluation

CONCLUSIONS

- Objective evaluation
- Personal conclusion
- Future work

Logistic Regression

This method involves summing the weights of the predictor variables, where each variable will have a weight between 0 and 1. The sum of all the variables will be our response variable. If this response variable has a probability greater than 50% of belonging to a particular group, it will be classified into that group; otherwise, it will be classified into the remaining group. To apply this method to our model, we have used the Scikit-learn library in Python

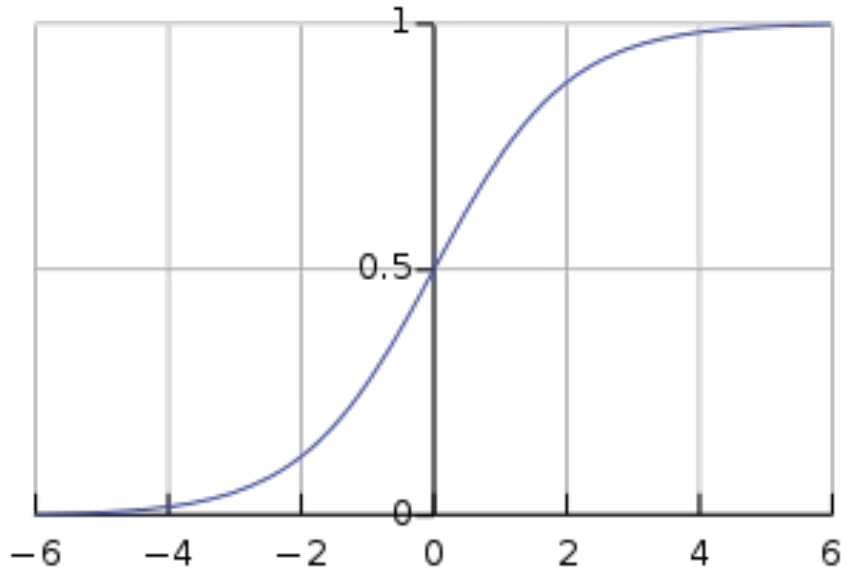
Logistic curve formula

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Where *t* is the value of the linear regression equiation:

$$t\,(xi) = \theta 0 + \theta 1x1 + \theta 2x2 + \cdots + \theta nxn$$

Logistic curve graph



Evaluation Methods

KEY CONCEPTS FOR THE EVALUATION

There are different evaluation metrics that will allow us to determine the percentage of successes of our model. There are 4 key concepts to keep in mind: training and testing, overfitting, random state and stratify

TRANING AND TESTING

It consists of dividing our data set into two subsets, one for training and one for testing. This division allows us to estimate how the model behaves when working with new data

OVERFITTING

This problem occurs when the model is too well trained. The model knows all the data with which it is trained and learns from the details and noise to adjust its predictions

RANDOM STATE

In order to be able to compare the performance of the different Machine Learning algorithms, we have decided to set a fixed value to divide the rows of the data set. This pseudo-random value is called random state

STRATIFY

Stratify allows us to have the same number of cases to predict in the training set and the test set

Evaluation Metrics

CONFUSION MATRIX

This matrix gives us information about how correct our predictions are. We will classify the results into 4 categories: True positives, false positives, and true negatives and false negatives

PRECISION y RECALL

Precision and Recall are two ratios that allow us to know the percentage of successes out of the total customer churns. Each metric has pros and cons

F1 - Score

This metric consists of a harmonic mean between Precision and Recall. Using the F1-Score simplifies our work so we don't have to balance between Precision and Recall metrics

ACCURACY

Accuracy is a ratio that allows us to know the percentage of successes over the total predictions

Evaluation Metrics

CONFUSION MATRIX

This matrix gives us information about how correct our predictions are. We will classify the results into 4 categories:
True positives, false positives, and true negatives and false negatives

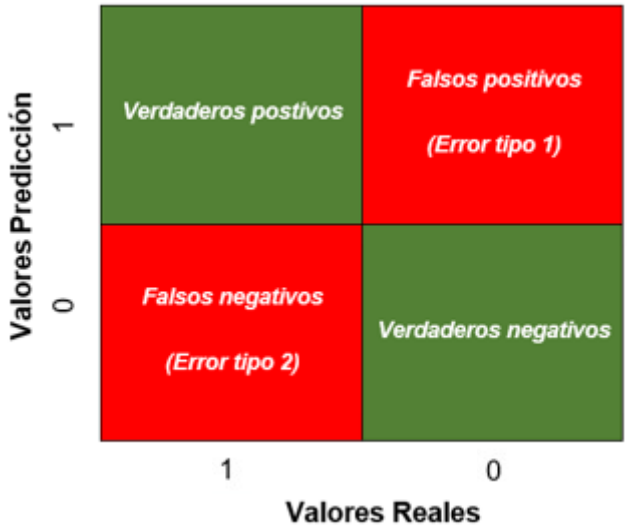


*Illustration names might be in Spanish since the project was originally documented in Spanish

Evaluation Metrics

PRECISION y RECALL

Precision and Recall are two ratios that allow us to know the percentage of successes out of the total customer churns. Each metric has pros and cons

Precision is a ratio that allows us to know the percentage of customers that we have identified as brand churn out of the total number of customers that are predicted to churn from the brand. Precision does not take into account type 2 error

$$Precision = \frac{Verdaderos\ positivos\ (VP)}{Verdaderos\ positivos\ (VP) + Falsos\ positivos\ (FP)}$$

Recall is a ratio that allows us to know the percentage of customers who leave the brand and we have correctly predicted the total number of customers who will leave the brand.
Recall does not take into account type 1 error

$$Recall = \frac{Verdaderos\ positivos\ (VP)}{Verdaderos\ positivos\ (VP) + Falsos\ negativos\ (FN)}$$

*Illustration names might be in Spanish since the project was originally documented in Spanish

Evaluation Metrics

F1 - Score

This metric consists of a harmonic mean between Precision and Recall. Using the F1-Score simplifies our work so we don't have to balance between Precision and Recall metrics

F1-Score formula

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Model evaluation in Jupyter Notebook

	precision	recall	f1-score
0.0	0.60	0.57	0.59
1.0	0.59	0.62	0.60
accuracy			0.60

Evaluation Metrics

ACCURACY

Accuracy is a ratio that allows us to know the percentage of successes over the total predictions

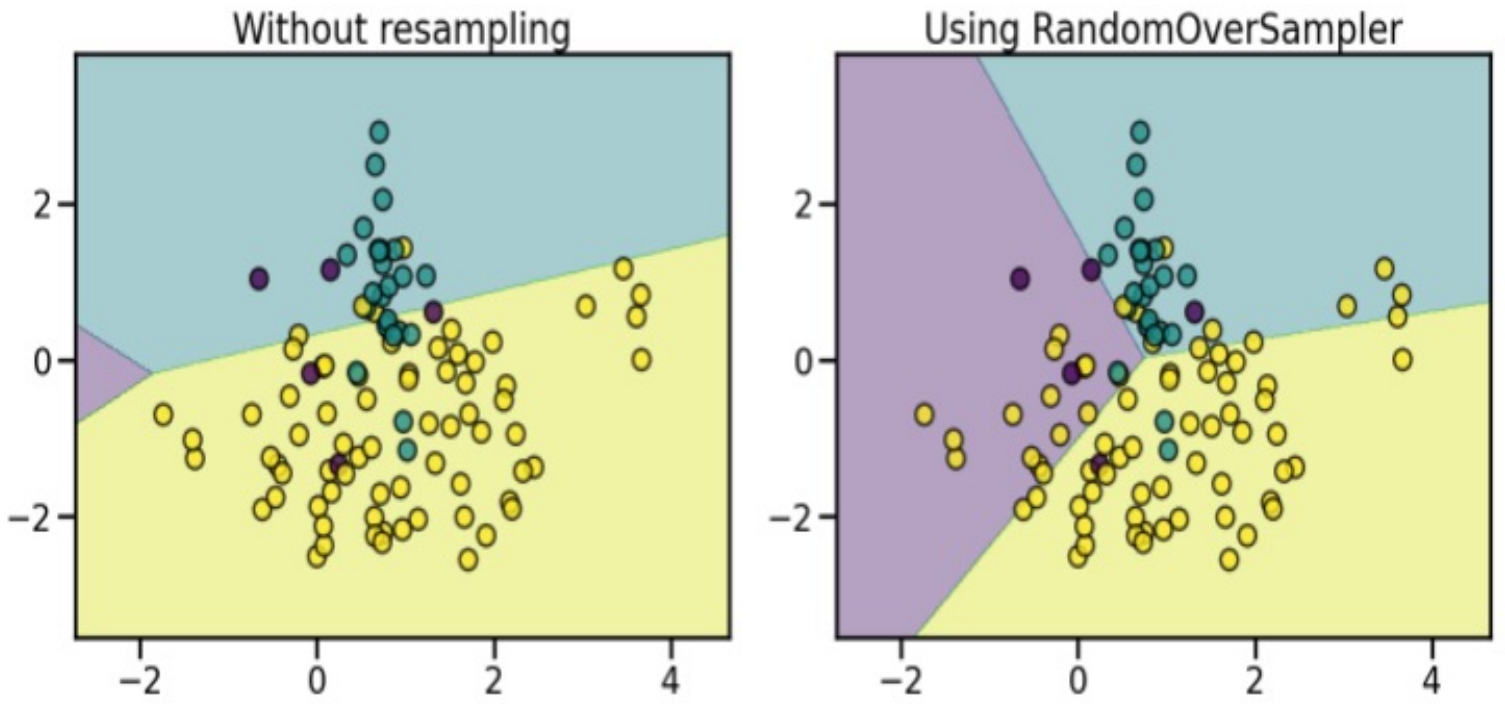
Accuracy formula

$$Accuracy = \frac{Verdaderos\ positivos + Verdaderos\ negativos}{Verdaderos\ positivos + Falsos\ positivos + Verdaderos\ negativos + Falsos\ negativos}$$

*Illustration names might be in Spanish since the project was originally documented in Spanish

Over-sampling

In our dataset, the churn rate is only 11.4%. Since we have little representation for this segment that we want to predict, our model will have a hard time predicting it. For this reason, it is necessary to use over-sampling techniques



Application and evaluation of the model

To apply our model, we will compare and evaluate the results for different values of the hyperparameter C. It will also be necessary to decide which is the most appropriate metric for the company's interests

Set C value

- A high value of C tells the model to give more weight to the training data. A lower value of C will indicate that the model should give more weight to complexity at the expense of fitting the data

- Therefore, a high value of hyperparameter C indicates that the training data is more important and reflective of real-world data, while a low value is just the opposite

Prioritize *Precision*, *Recall* or F1-Score

- Depending on the brand's strategy, certain metrics will be prioritized, leading to the determination of a specific value for C

	0 (no churn)			1 (churn)		
C	Precision	Recall	F1 - Score	**Precision**	**Recall**	**F1 - Score**
0.000001	0,60	0,45	0,52	0,56	0,69	0,62
0.00001	0,60	0,46	0,52	0,56	**0,70**	0,62
0.0001	0,62	0,50	0,55	0,58	0,69	**0,63**
0.001	0,61	0,58	0,59	**0,60**	0,63	0,61
0.01	0,60	0,57	0,59	0,59	0,62	0,60
0.1	0,60	0,59	0,59	**0,60**	0,62	0,61
1	0,60	0,57	0,59	0,59	0,62	0,60
10	0,60	0,59	0,59	**0,60**	0,61	0,61
100	0,60	0,59	0,59	**0,60**	0,61	0,61
1.000	0,60	0,59	0,59	**0,60**	0,61	0,61
10.000	0,60	0,59	0,59	**0,60**	0,61	0,61

KEYNOTE STRUCTURE

DATA STUDY AND UNDERSTANDING

- Data Structure
- Data Extraction
- Data Preparation

DATA ANALYSIS AND FEATURE SELECTION

- Customer profiling
- Correlations between variables

MODELING

- Logistic Regression
- Evaluation methods
- Evaluation metrics
- Over-Sampling
- Model application and evaluation

CONCLUSIONS

- Objective evaluation
- Personal conclusion
- Future work

Project Conclusions

Objective evaluation

- The CRISP-DM methodology has been successfully applied
- Data analysis has represented a very high percentage of the time dedicated to this work.
- A model has been achieved that predicts customer churn with up to 70% accuracy

Personal conclusion

- Multiple possibilities of Data Mining and Machine Learning
- Difficulties in working with a real unstructured database

Future work

- Possibility of deepening the analysis at a higher level than the current one
- Possibility of using another subset of data
- Possibility of using other prediction methods other than logistic regression
- Possibility of using another over-sampling method