Week 9 deliverables

**Data Analyst: Cross selling recommendation**

1. **Team member's details**

**Group Name: Dreamers**

Name:  Surabhi mahawar

Email: [surabhimahawar@gmail.com](mailto:surabhimahawar@gmail.com)

Country:  India

College/Company:  APJ Kalam Technical University

Specialization ( Data Science, NLP, Data Analyst): Data analyst

Name:  Deborah Adeyemi

Email: [adeyemianuoluwapod@gmail.com](mailto:surabhimahawar@gmail.com)  ,

Country:  United Kingdom,

College/Company:  University of the West of England(UWE)

Specialization ( Data Science, NLP, Data Analyst): Data analyst

Name:  Kseniia Nosenko

Email: [kseniianosenko@gmail.com](mailto:kseniianosenko@gmail.com)

Country:  Germany

College/Company:  Higher School of Economics

Specialization ( Data Science, NLP, Data Analyst): Data analyst

Name:  Pranav Walia

Email: waliap@miamioh.edu

Country:  America

College/Company:  Miami University

Specialization ( Data Science, NLP, Data Analyst): Data analyst

**Problem description**

XYZ Credit union in Latin America is performing very well in selling Banking products (Credit card , deposit amount, retirement account, safe deposit box), but their existing customer is not buying more  than 1 product which means bank is not performing good in cross selling (Banks is not able to sell their other offering to existing customers). XYZ credit union decided to approach ABC analytics to solve their problem. 

**GitHub Repo Link**

<https://github.com/pranav611/Final-Project>

**Final Recommendation**

The recommendation proposed will be developed for the “Affluent Low Income Risk” market. We will focus on improved cross-selling techniques at the time of sale of any product by the bank. This specific market will be highly price sensitive to the offers, therefore, the lower the price, the more cross selling will be done.

1. **Data understanding**

There are 2 files: Test and Train. Test file hasn't any info about products and we can’t use this file for analysis.

Let’s check the number of rows in train datasets, what type of variables are there and whether values are null (info in %).

|  |  |
| --- | --- |
| Table  Description automatically generated | Table  Description automatically generated |

Train dataset contains info about customers and their products.

Train dataset contains 48 features and 13 647 309 rows. There are no duplicates.

Some of features was defined as object, but it is Numerical Variables:  age, antiguedad.

Some features in opposite were defined as float64, but it is categorical features: ind\_nuevo, indrel, tipodom, cod\_prov, ind\_actividad\_cliente

In general:

Continuous Variables: 'age', 'antiguedad', 'renta'

Categorical Variables: others

There are some features with the same number of missing values, I expect those relate to the same rows. We delete rows with a lot of missing values (0.2% of data).

Null values after deleting rows with a lot of missing values:

Table

Description automatically generated

**Describe data:**

* **numeric**

Table

Description automatically generated

**Outliers in numeric data:**

The feature 'age' has some rows with customers  who are older than 100 years and a few customers who are very young. We think there is a lot of incorrect data. We can see that customers who are older 20 and younger 90  are most. One of the ways to overcome outliers is to unite the youngest and most adult customers into groups.

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Graphs after corrections:

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

The feature 'antiguedad' (Customer seniority (in months)) contains 38 rows with value -99999. It makes the data very skewed. We should delete rows with this value, because it is an unknown value:

A picture containing graphical user interface

Description automatically generated

The box plot feature’s 'antiguedad' after deleting:

Chart, box and whisker chart

Description automatically generated

The feature Renta is also very shifted, because there is 18,9% data much more than 75% quartile.

Chart

Description automatically generated

Almost 19% of data are the outliers. And 20% of data is null. Delete these rows will be incorrect. It is necessary to carry out work on the replacement of zero values and emissions. For NA values it can be for example mean/median/mode/segmented approach etc.

For outliers it can be grouping.

There is no correlation between numerical features:

Chart

Description automatically generated

* **categorical**

1. ind\_empleado - Employee index. No NA values, 99 % rows have the value N not employee.
2. pais\_residencia - Customer's Country residence. No NA values. 118 unique values. 99 % rows have the value ES**.**
3. sexo - Customer's sex. A small number NA values (0.0005% or 70 rows). It is necessary to carry out work on the replacement of NA values (the most popular values for example).
4. fecha\_alta - The date in which the customer became the first holder of a contract in the bank. No NA values.
5. ind\_nuevo - New customer Index. 1 if the customer registered in the last 6 months. Data type float, should be changed to categorical. No NA values.
6. indrel - 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month). No NA values. If you build an ML model, it could be better to change 99 on 0 because it is scaled for ML models.
7. ult\_fec\_cli\_1t- Last date as primary customer (if he isn't at the end of the month) and conyuemp- Spouse index. (1 if the customer is spouse of an employee). have 99% null values. According to the instructions  conyuemp feature should contain number 1 if the customer is spouse of an employee. In dataset the feature conyuempcontain (N, S, nan) values. I suppose that N=No, S = Si (Yes). The number of clients with value ‘S’ = 17. *We can delete these features because they contain too small info for analysis.*
8. indrel\_1mes - Customer type at the beginning of the month and tiprel\_1mes- Customer relation type at the beginning of the month. There are 0.89% NA values. It is necessary to carry out work on the replacement of NA values.
9. indresi - Residence index. No NA values. 99% of customers have the same residence country as  the bank country.
10. indext - Foreigner index. No NA values. (S (Yes) or N (No) if the customer's birth country is different than the bank country) 0.95 % of rows have value N.
11. canal\_entrada - channel used by the customer to join. 162 unique values. The most popular is KHE. 1.16% are NA values. We can replace the missing values with the most popular values in general or by region or something else.
12. indfall - Deceased index - 0.99 of rows have value N (not). No NA values.
13. tipodom - Address type. 1, primary address.  No NA values.
14. cod\_prov(Province code ) and nomprov(Province name) explain the same thing. I suppose delete feature  cod\_provand change on categorical values feature nomprov. They have the same number of NA values - 0.48%.
15. ind\_actividad\_cliente - Activity index (1, active customer; 0, inactive customer). No NA values. There are 54% inactive customers.
16. segmento - segmentation: 01 - VIP, 02 - Individuals 03 - college graduated. 1.18% NA values. PARTICULARES customers are 58%, UNIVERSITARIO customers are 36%, TOP customers are 4%
17. Other features describe the product and customer's product availability.

**Approaches to overcome problems like NA value, outlier etc:**

**NA values:**

* replacing with a mean value where there are not many missing values, this will not affect the result due to the small volume.
* replacing the average value based on the data of another feature (Segmento-Renta).
* Alternatively, you can set a constant value for NA-marked values. For example, you can put in a special string or numerical value

**Outliers:**

* Grouping data (all clients with an age of less than 20 years, set the age of 19 years to those from 90 - 91 years)
* Moving from numeric data to categorical data (renta)