bip.

HERE TO DARE

Exemption Code Mapper

Agenda

1. Introduction
2. Preprocessing
3. Model Deployment
4. User Interface
5. MLOps
6. Discussion

HERE TO DARE

1 Introduction

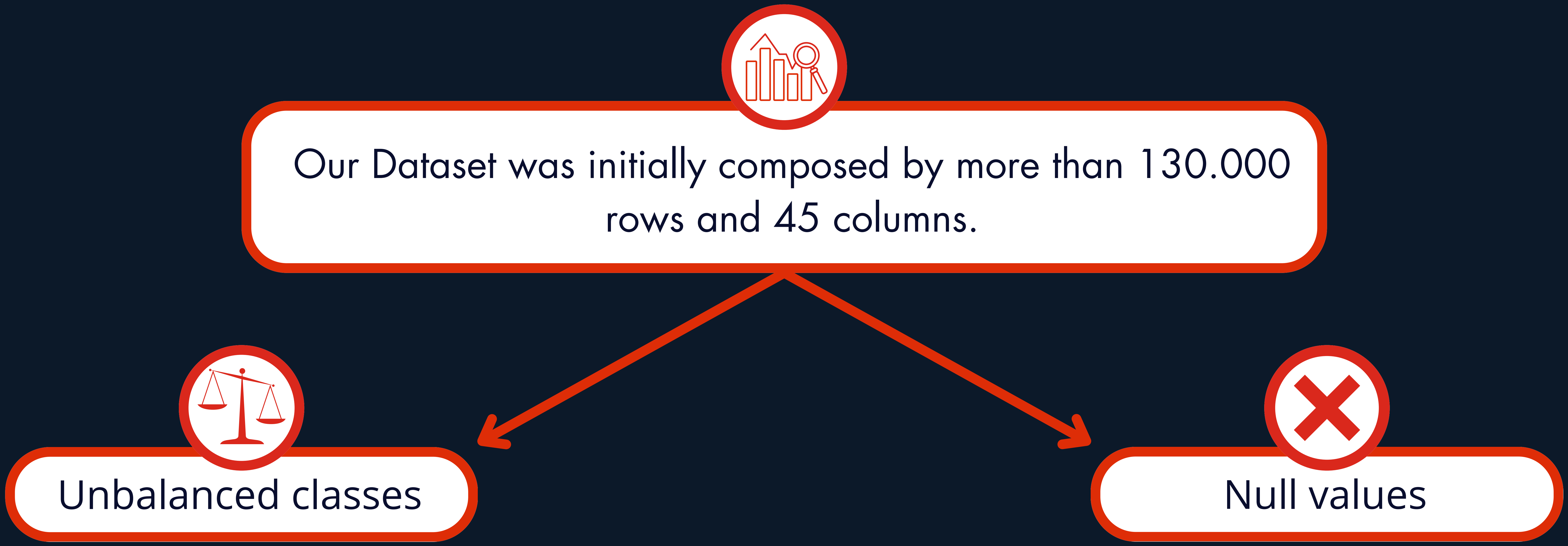
Introduction & Our goal

The objective is to create a model that predicts the **exemption VAT** code for invoices.

To achieve the objective, we were provided with a dataset concerning invoices and their associated characteristics.

2 Preprocessing

Our data



Our Dataset was initially composed by more than 130.000 rows and 45 columns.

Unbalanced classes

Null values

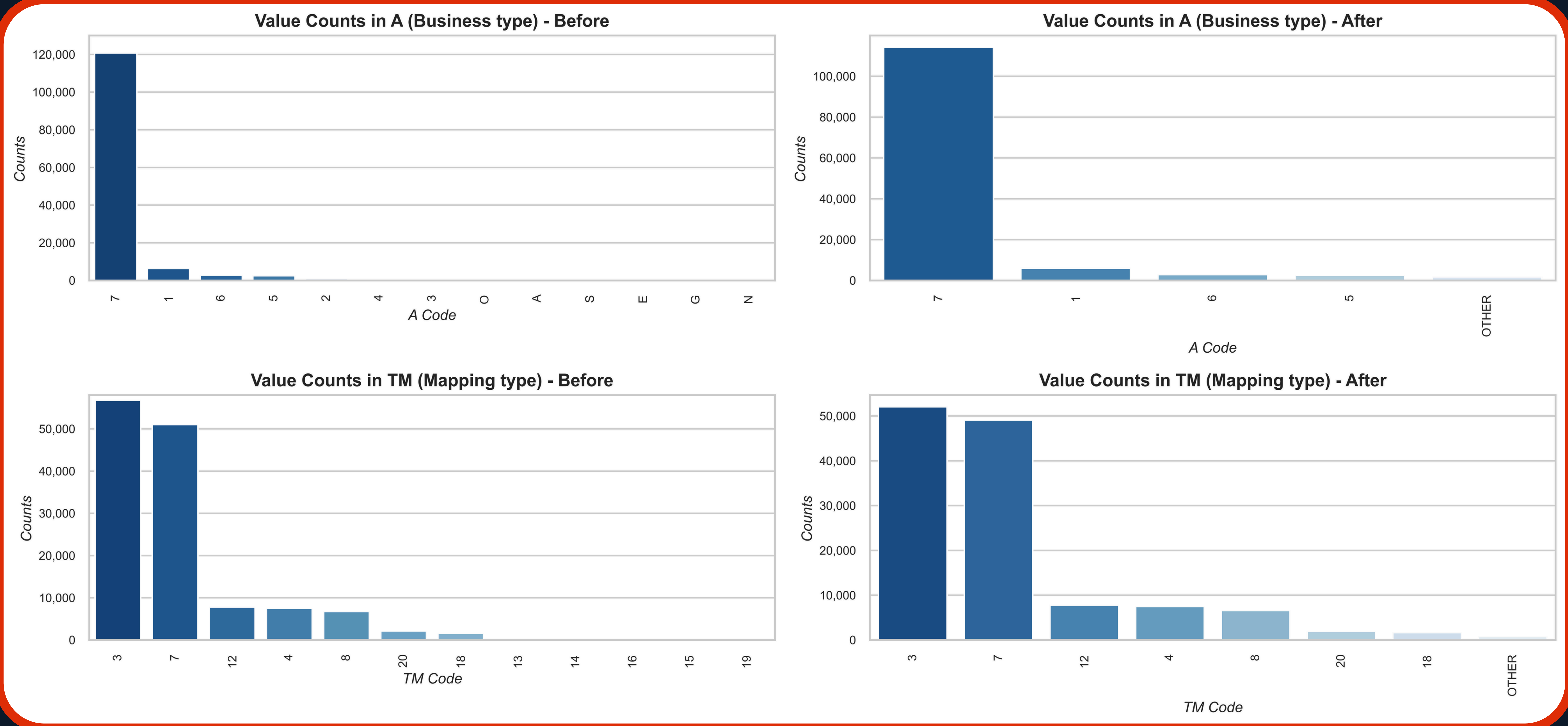
Null values

1. Initially, there were numerous columns with over half of their values null.

2. We dropped columns with more than 60% null values.

3. The remaining NaN values were filled with the most frequent class within the variable.

Unbalanced Classes (1/2)



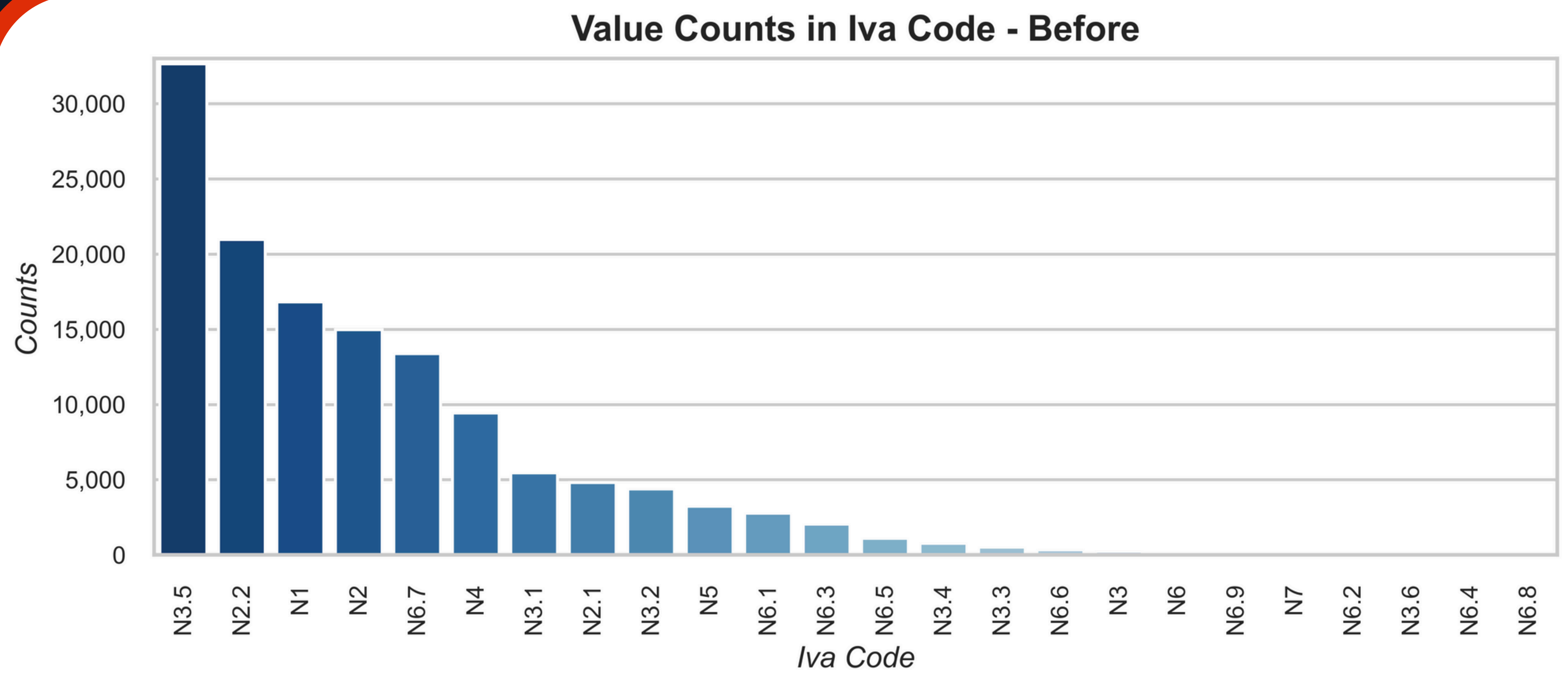
Value Counts in A (Business type) - Before

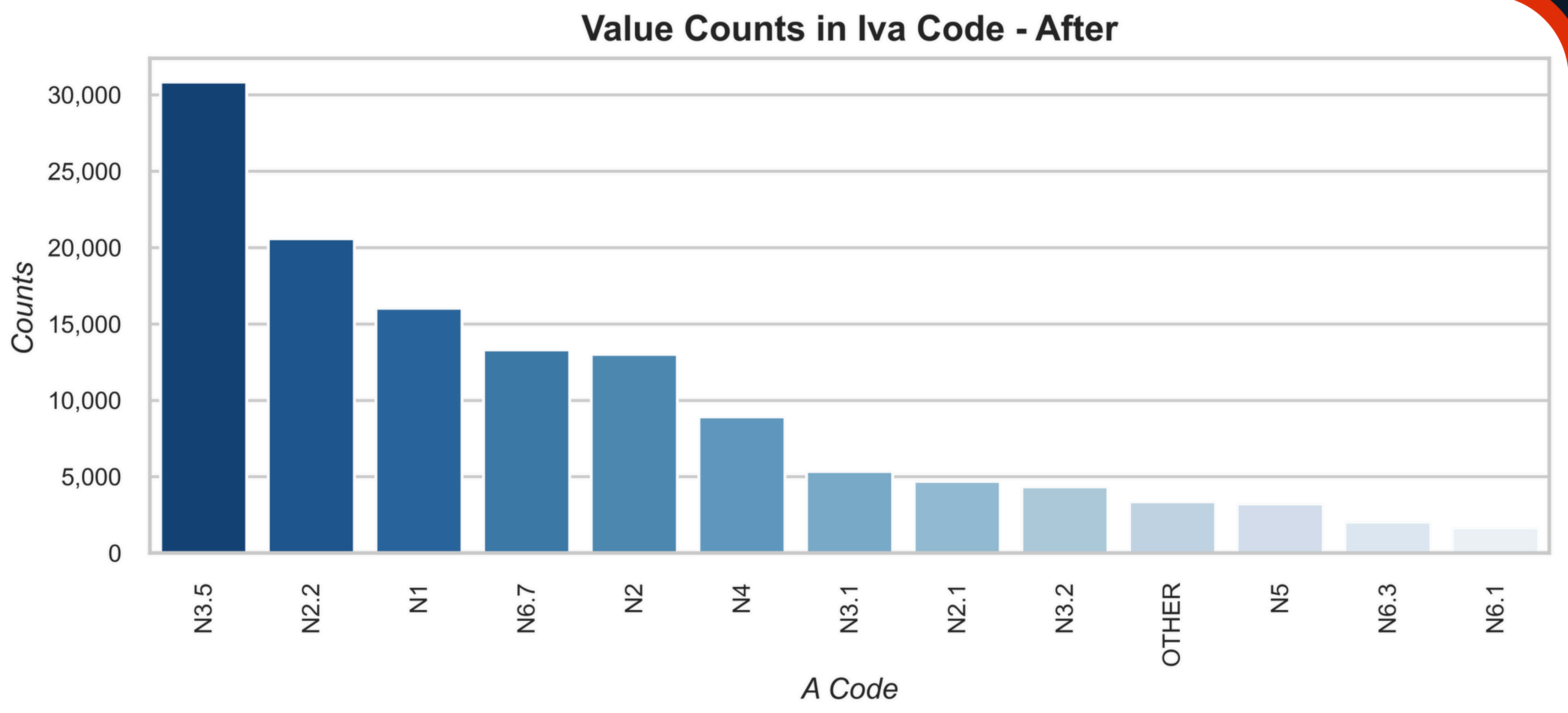
Value Counts in A (Business type) - After

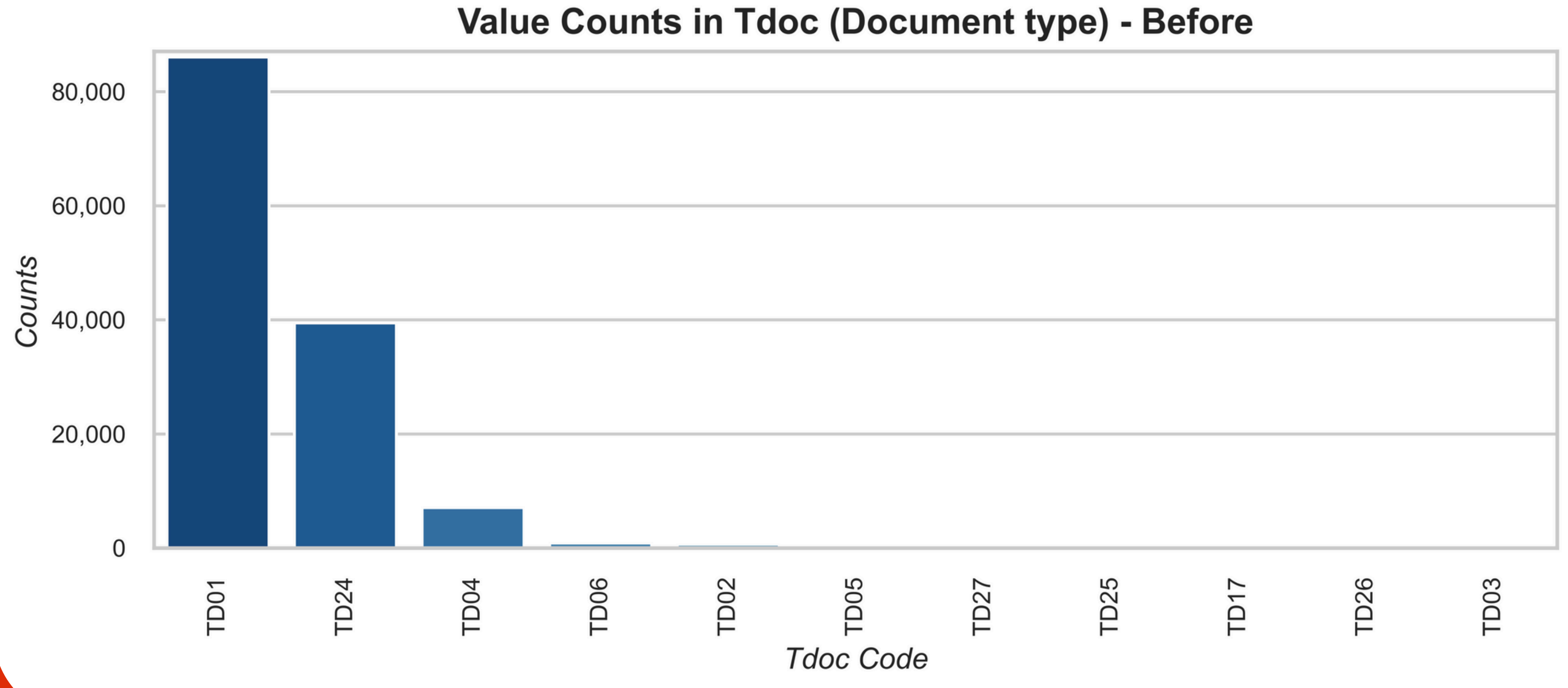
Value Counts in TM (Mapping type) - Before

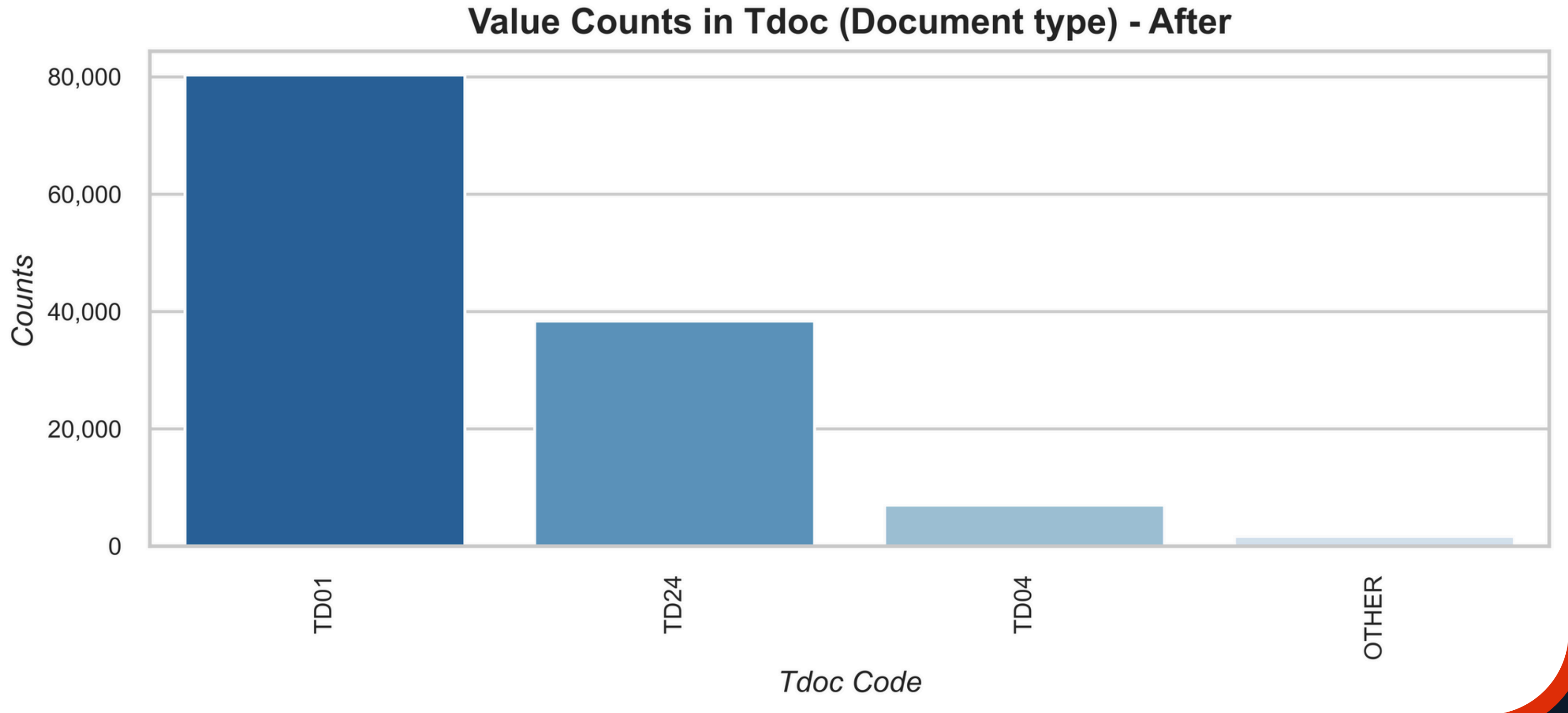
Value Counts in TM (Mapping type) - After

Unbalanced Classes (2/2)

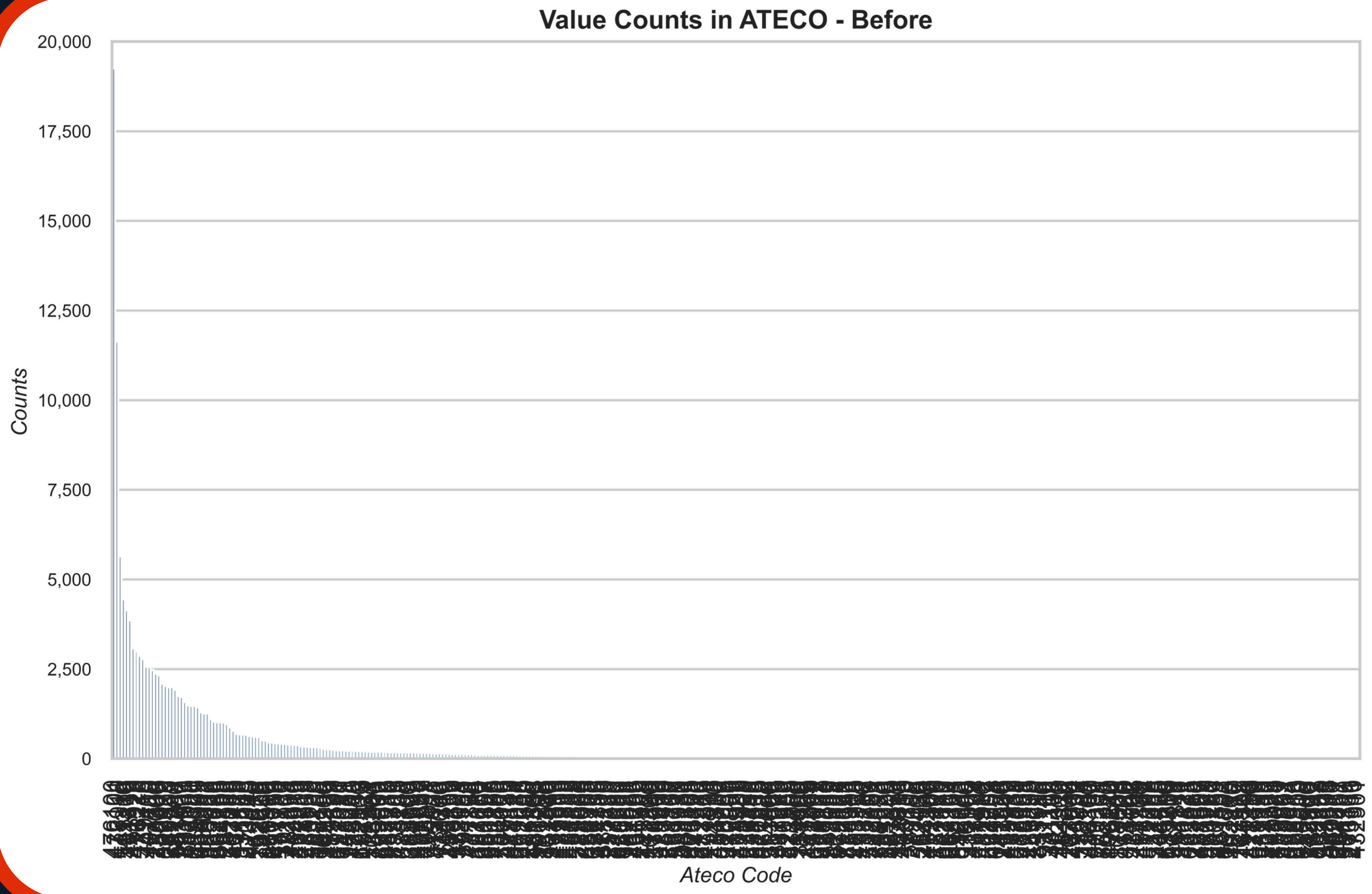

Value Counts in Iva Code - Before

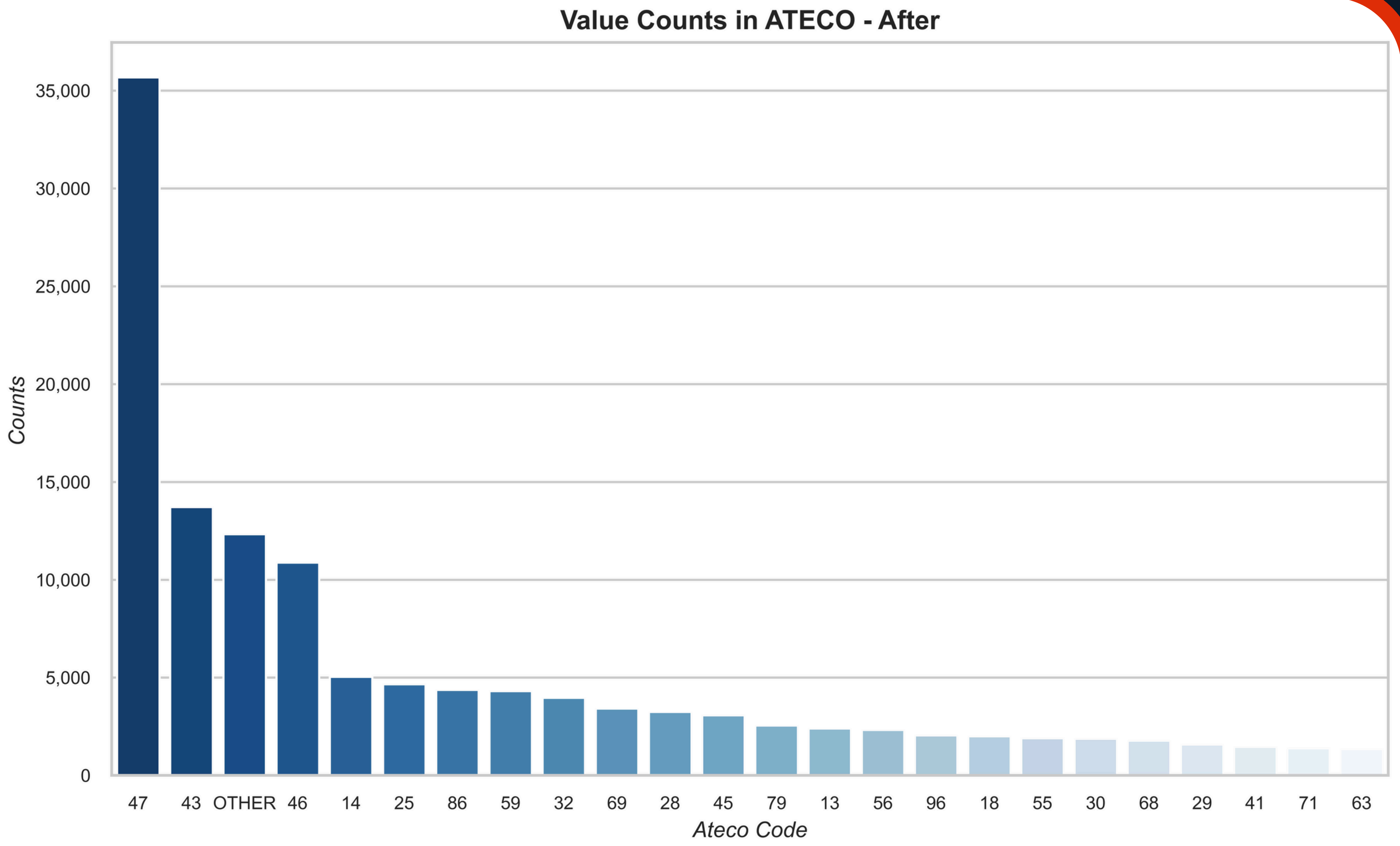

Value Counts in Iva Code - After


Value Counts in Tdoc (Document type) - Before


Value Counts in Tdoc (Document type) - After

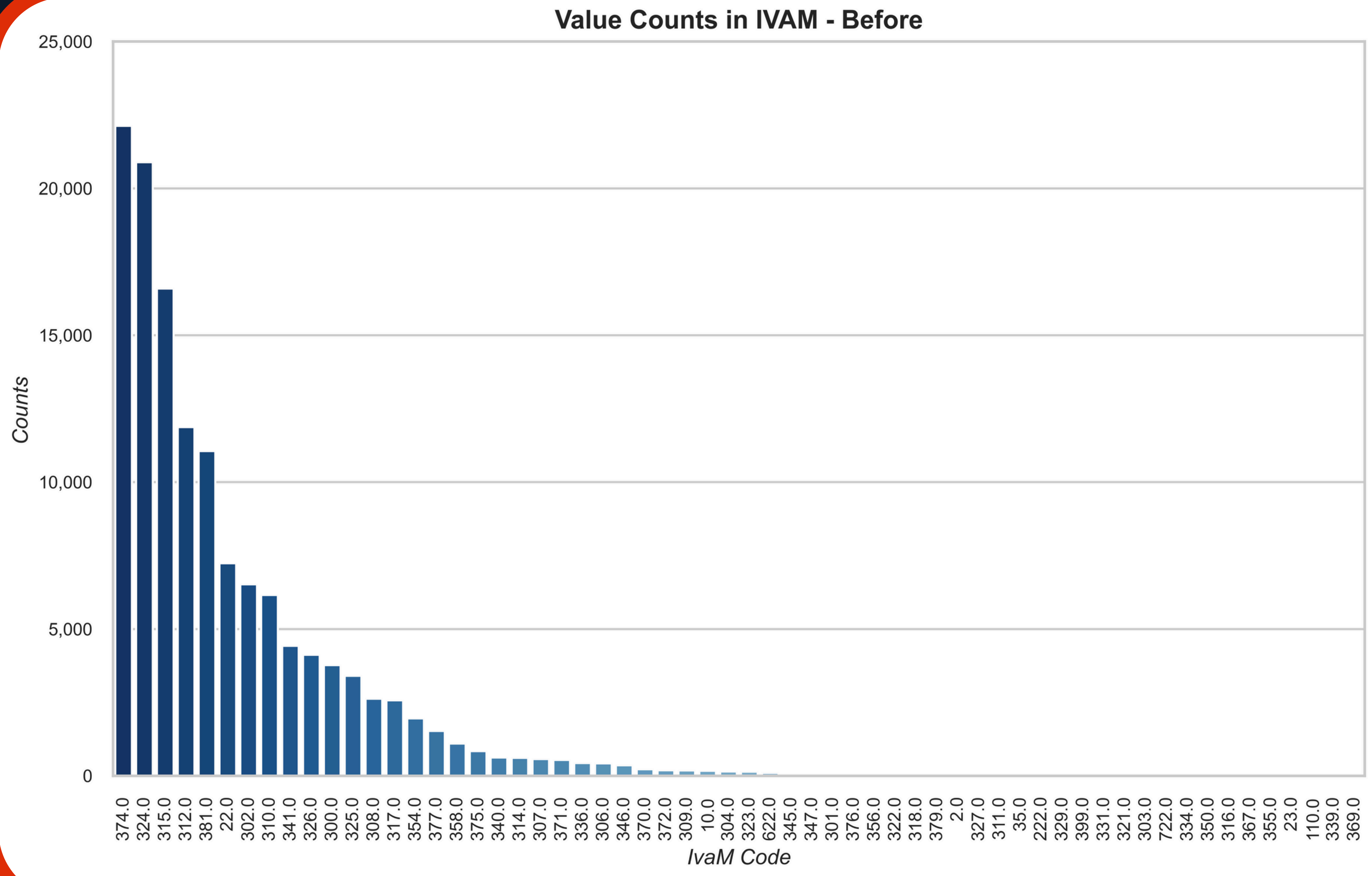
Focus on Ateco

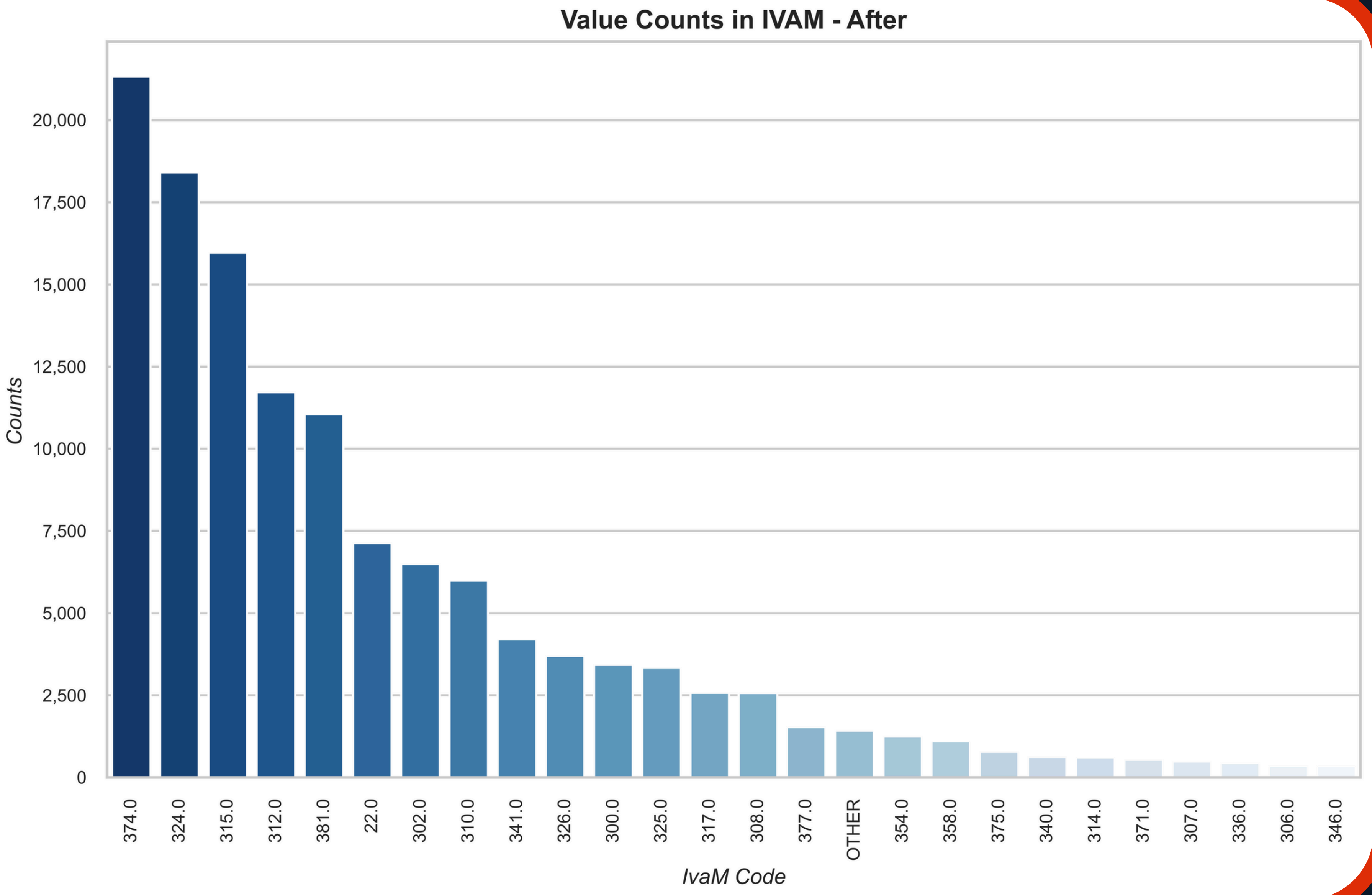

Value Counts in ATECO - Before


Value Counts in ATECO - After

For the ATECO code, our approach involves two steps: first, extracting the initial two digits of each code, and then consolidating classes with limited observations into an 'others' category.

Focus on IvaM

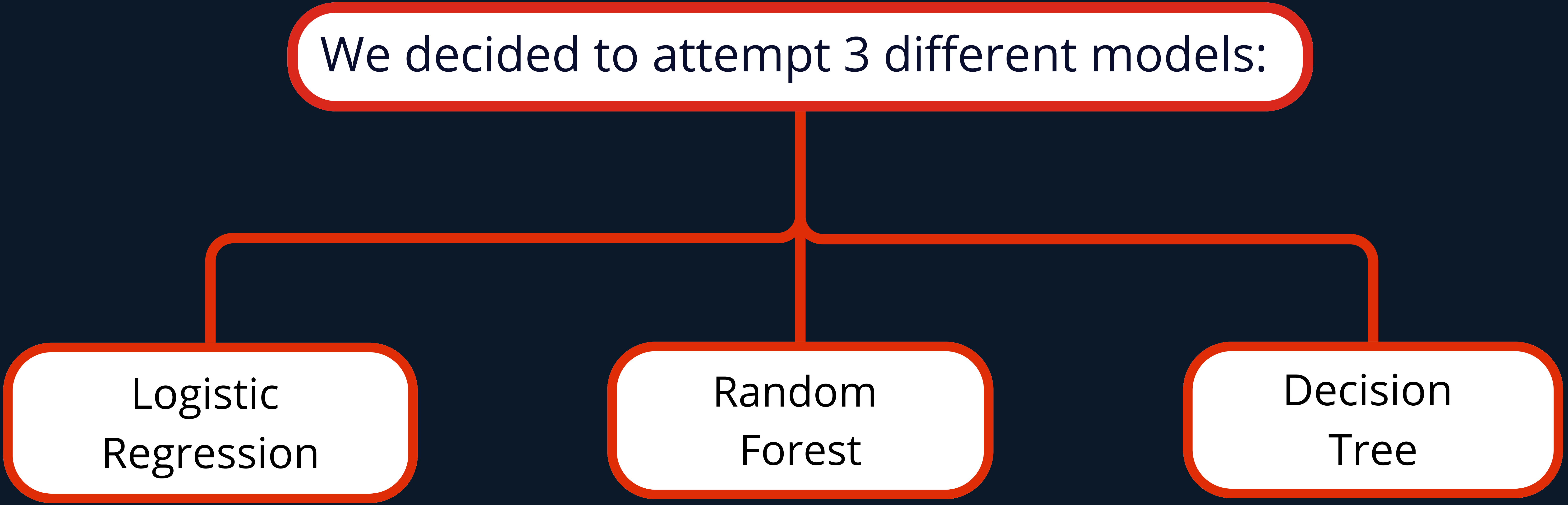




For our response variable, we opted to group classes with fewer observations, employing a lower threshold of 250 to maintain high sensitivity in the model.

3 Model Deployment

Our approach



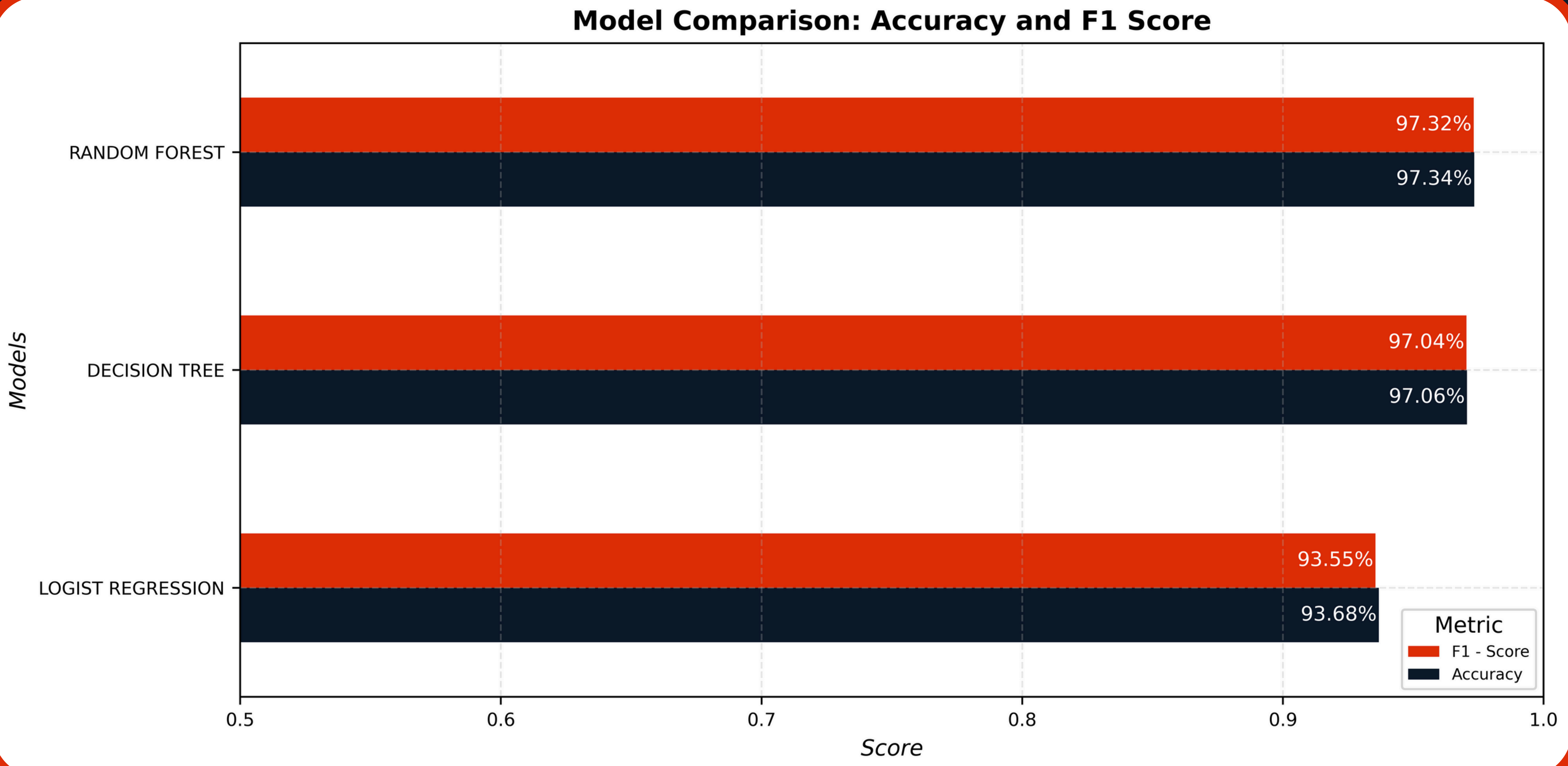
We decided to attempt 3 different models:

Logistic Regression

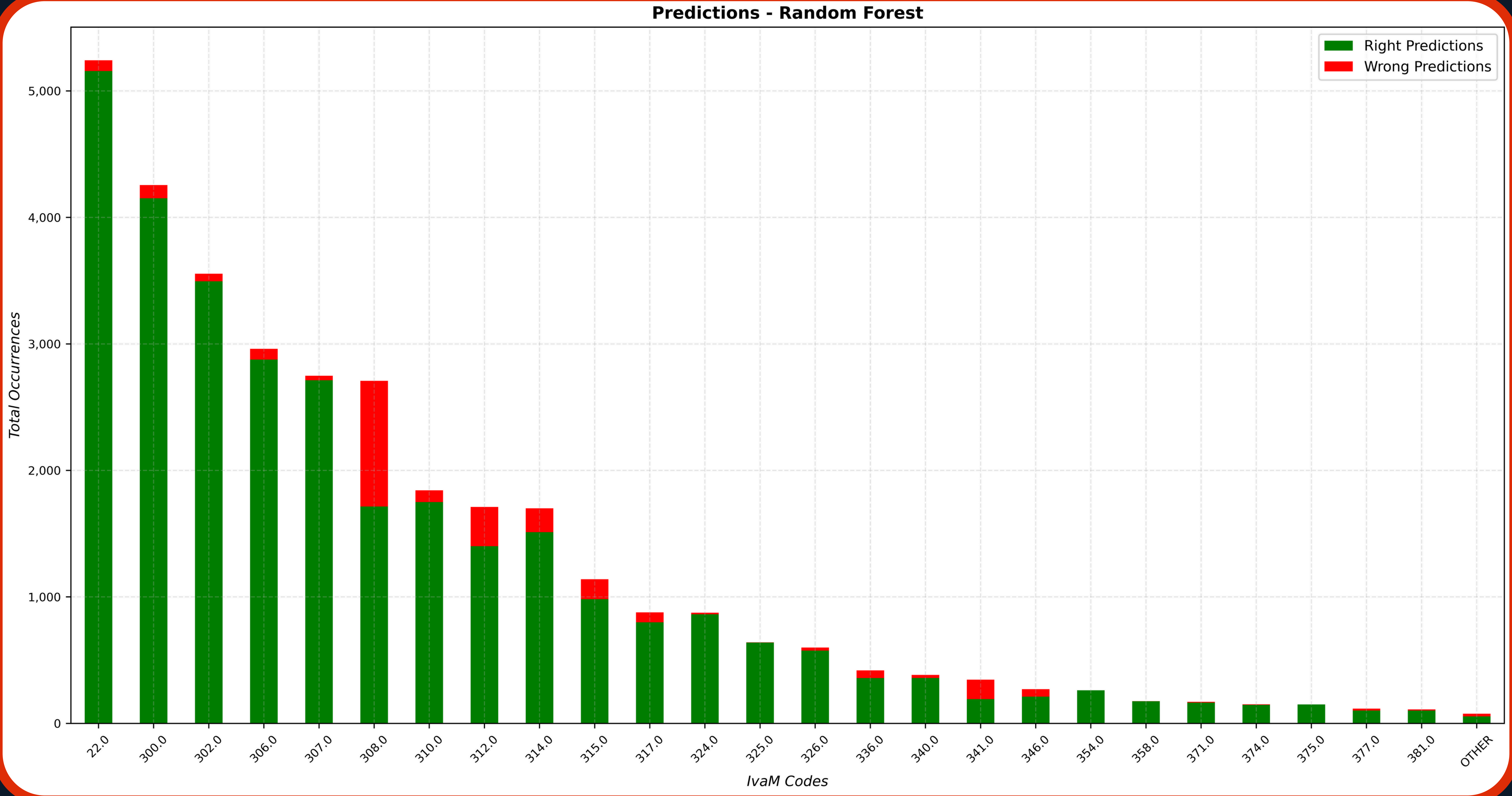
Random Forest

Decision Tree

Model comparison



Final approach - Random Forest



HERE TO DARE

4 User Interface

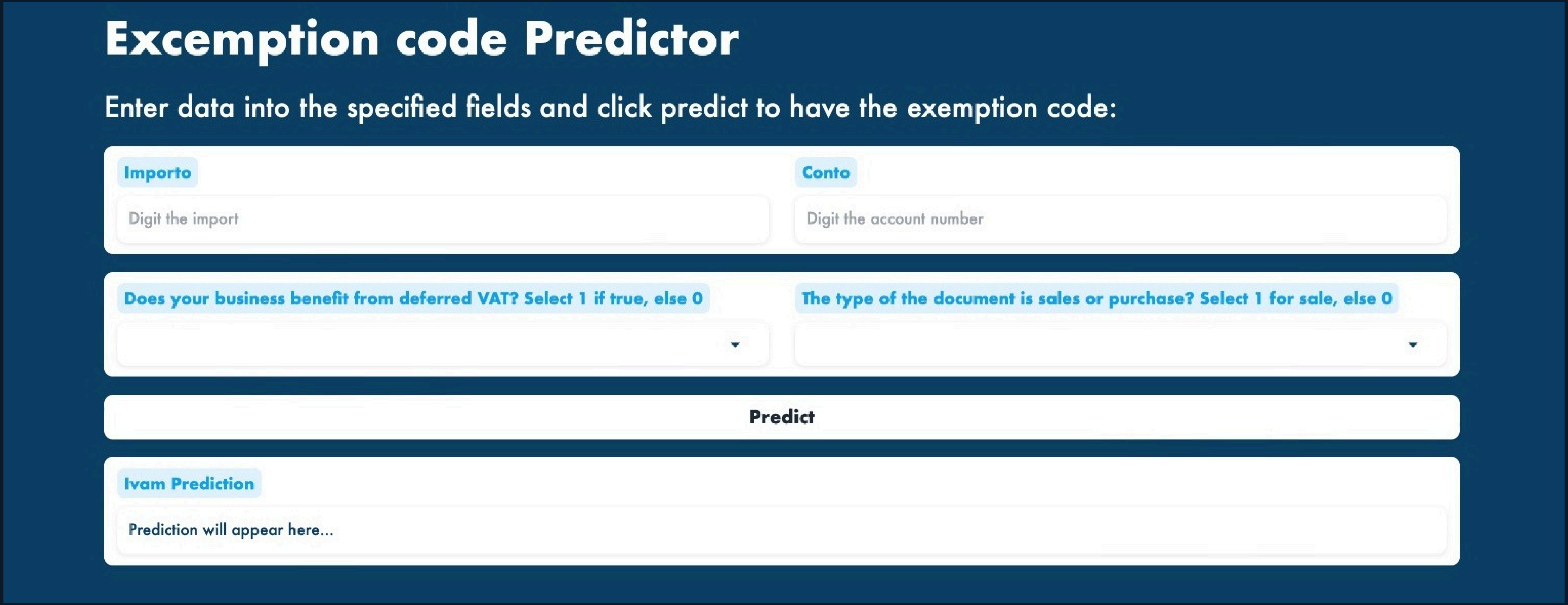
User interface

Trade-off between precision of the model and usability of the interface

Performed feature selection to to attain the optimal model, leveraging only four features.

http://127.0.0.1:7867

User interface



Excemption code Predictor

Enter data into the specified fields and click predict to have the exemption code:

Importo

Digit the import

Conto

Digit the account number

Does your business benefit from deferred VAT? Select 1 if true, else 0

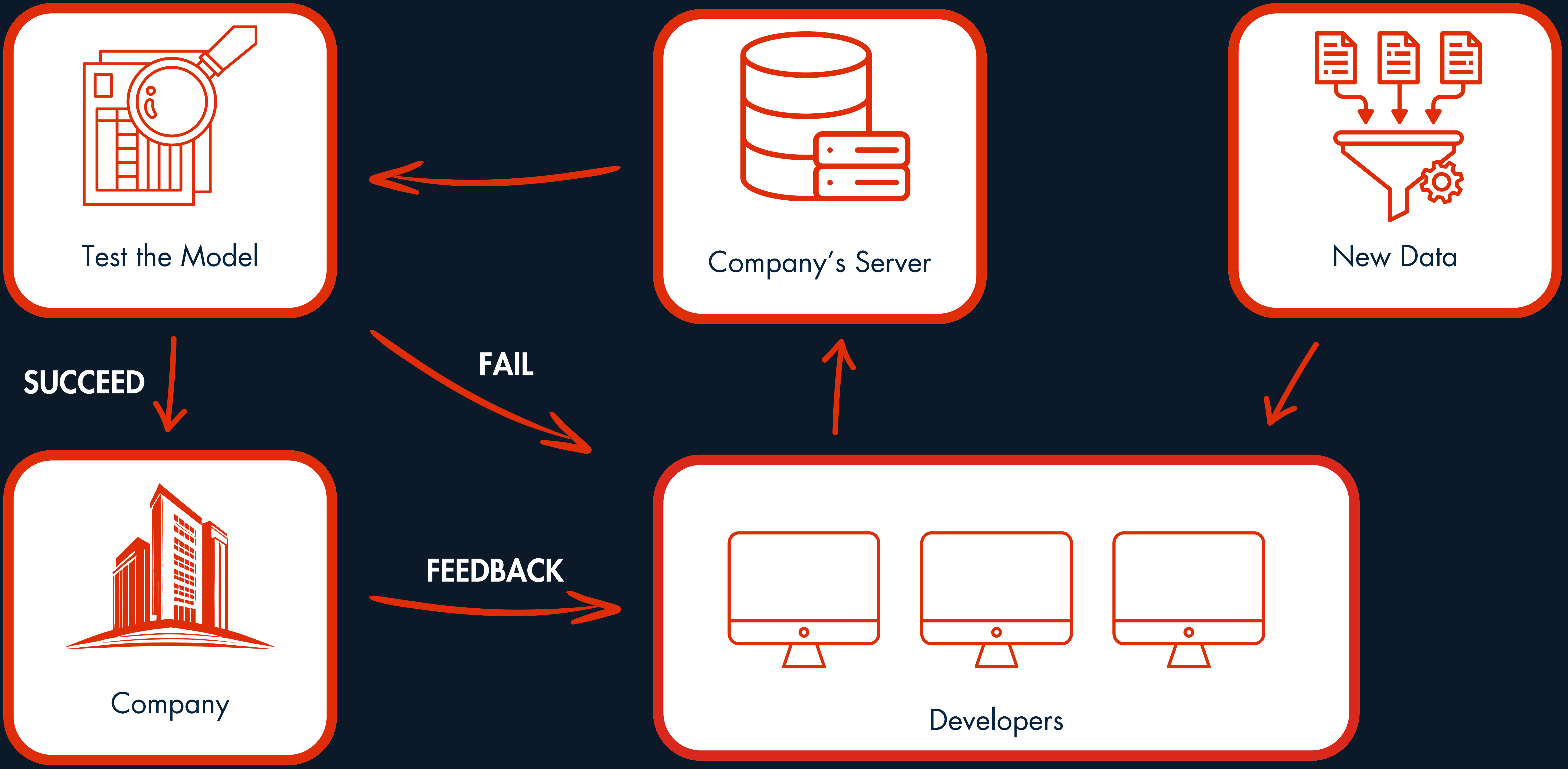
The type of the document is sales or purchase? Select 1 for sale, else 0

Ivam Prediction

Prediction will appear here...

5 MLOps

MLOps - Machine Learning supply chain



MLOps - Our Proposal

Data pipeline

1. Automation of data collection.
2. Automation of data preprocessing.
3. Implementation of new data to continuously train the model.

6 Discussion

Discussion

Overall we are satisfied with the performance of our model. But we are aware of its inability to predict the exemption code included into the class "Others", however we would be able to fix this by using a more balanced dataset.

bip.

HERE TO DARE

Thank You!

for the Attention

Vincenzo Camerlengo	773731
Daniele De Robertis	787291
Raffaele Torelli	775831