Predicting Health Insurance – Checkpoint 1

Group G

Daniel Carneiro, up202108832
Gonçalo Costa, up202108814
Eduardo Oliveira, up202108690

Index

- Loading Data / Data Exploration
- Data Preprocessing
- Models / Submission File
- Difficulties
- Future Developments

Loading Data / Data Exploration

- 15 Columns (But 3 are not necessary: '*Unnamed: 0*', '*custid*', '*code_column*')
- Zero duplicated rows
- 1686 Missing Values in '*housing_type*', '*num_vehicles*', '*gas_usage*' and '*recent_move_b*' (25515 in '*is_employed*' are not considerate missing values)

Data Preprocessing

- Dropped the unnecessary columns (‘*Unnamed: 0*’, ‘*custid*’, ‘code_column’)
- Handling with missing values (‘*housing_type’ -> Unknown*, ‘*num_vehicles*’ -> 0 , ‘*gas_usage*’ -> 0 and ‘*recent_move_b*’ -> *Unknown*)
- Handling Outliers
  - *‘gas_usage’* : (we did a log)
  - *‘income’* : (we divide by 12 for put all values in month scale)

Models / Submission File

- Before we train the model:
  - Label encoding and Binarization
  - Did some tests between **SMOTE** and **Oversampling**

- Training models:
  - 90% for train and 10% for test
  - KNN ; Decision Tree ; SVC ; Naïve Bayes ; Neural Networks ;

	f1	accuracy	recall	auc
Decision Tree	0.896346	0.896341	0.896341	0.896385
KNN	0.854834	0.855183	0.855183	0.918634
SVC	0.663046	0.665396	0.665396	0.714960
MLP	0.658865	0.664634	0.664634	0.638087
Gaussian Naive Bayes	0.583073	0.606707	0.606707	0.690732

- Submission File:
  - Best result until now was **0.70168**:

1	**Gonçalo Bessa Costa**		0.70168	4	1d

Difficulties

- Income outliers
- Missing Values in '*housing_type*', '*num_vehicles*', '*gas_usage*' and '*recent_move_b*'

Future Developments

- Improve the models
- Test different ways to deal with outliers and Missing values
- Add more columns with a good relation
- Basically, have a better score in submission