The Dataset

The training set (housing_data_train.csv) is composed by 16762 samples, each sample contains information about a listing on a housing platform. Goal of the classification is distinguishing between scam and legit listings at creation time, where the majority of these features are actually available. A brief explanation of the available features is provided in the subsequent table.

Feature Name	Description
LISTING_KIND	0 - entire place; 1 - private room; 2 - shared room
LISTING_CITY	The city where the listing is located
LISTING_PRICE	The monthly rent (€) of the listing
IS_ARCHIVED	If the advertiser creating the listing has been archived or not.
ARCHIVE_REASON	The reason why the advertiser creating the listing has been (possibly) archived.
LOGIN_COUNTRY_CODE	The (last) country where the advertiser logged in from.
LISTING_COUNTRY_CODE	The country where the listing is located.
LISTING_REGISTRATION_POSSIBLE	If it’s possible to use listing’s address for registering at the city's municipality.
ADVERTISER_COMPLETENESS_SCORE	Percentage of completeness of the advertiser’s profile.
MANAGED_ACCOUNT	If the advertiser creating the listing is managed by our employees or not.
HAS_PROFILE_PIC	If the advertiser creating the listing has a profile pic or not.
BROWSER	The browser used to create the listing.
OS	The operating system used to create the listing.
ANONYMISED_EMAIL	The email address (anonymised) of the advertiser. Letters have been changed with random letters, numbers have been changed with random numbers, all the other characters have been maintained. The email domain has been maintained.
IS_SCAMMER	Whether the listing is a scam (1) or not (0).

Feature Engineering - Sampling

The given “housing_data_test.csv” set has less columns than the “housing_data_train.csv”, so the columns “IS_ARCHIVED” and “ARCHIVE_REASON” are dropped in order to create a train and a test set that have the same columns.

Also, the boolean “MANAGED_ACCOUNT” column is converted to integer (True -> 1, False -> 0). We decide to keep this column and not to drop it because we think that if the advertiser creating the listing is managed by our employees, then the listing is more credible, meaning that it is less possible to be a scam.

Then the problem of “object” type columns has to be tackled. These columns have to be converted to numerical data in order to be used or to be completely dropped.

The “BROWSER”, “OS” and “ANONYMISED_EMAIL” columns have many “nan” values in both train and test sets. Also, we don’t see how these features could improve the prediction accuracy of scams, so they were dropped.

Then, the only remaining “object” type columns are “LISTING_CITY”, “LOGIN_COUNTRY_CODE” and “LISTING_COUNTRY_CODE”. Considering them as useful features, they are not dropped and are converted into integers by mapping each country and city code to an integer number.

Finally, we notice that the “LOGIN_COUNTRY_CODE” has many “nan” values. Also, this column and the “LISTING_COUNTRY_CODE” has the same country codes. So, the missing values of the first feature are filled by copying the value of the second for the same listing. We do that based on the fact that in the most cases these two features have the same value. Furthermore, we make the assumption that the reason for the missing values of the “LOGIN_COUNTRY_CODE” is the “obvious” fact that it is more common to advertise a listing which is located at the same country as you do.

Algorithms

In this project, the performance of 5 common classification algorithms is tested: Decision Trees, Random Forest, KNN (K-Nearest Neighbor), SVM (Support Vector Machines) and Logistic Regression.

Evaluation and Metric

In order to determine which algorithm performs the best the train set is splitted into a new train and test set (70% - 30%). So, 5 models are implemented.

At first, the metrics that are used in order to evaluate the performance of every model are the following: accuracy score, average precision score, f1 score and recall score. However, it is obvious that the dataset is “unbalanced” as the “not scam” listings are far more than the “scam” (approximately 95% to 5%). Consequently, the most suitable metric to our case is the f1 score and finally this is used for model evaluation.

Results

Fortunately, in this project the labeled test is available, so the predictions made by our 5 models are evaluated on the real data. As we can see at the next array, which presents the results for the five algorithms, the algorithm that performs the best in terms of **f1 score** is the **Random Forest**.

Decision Trees	Random Forest	KNN	SVM	Logistic Regression
0.5394736842105263	0.6464646464646465	0.45116279069767445	0.3137254901960785	0.0