# Predicting Risk of Long-term Unemployment

## PROJECT OVERVIEW

Câmara Municipal de Cascais (CM-Cascais) is the local government of the municipality of Cascais. It is one of most livable and touristic cities in Portugal, and is home to over 200,000 residents. The municipality is increasingly using data to inform decisions.

The project, part of Data Science for Social Good Europe, aims to support CM-Cascais in understanding patterns of unemployment in the municipality and to develop a prediction system to identify individuals at the higher risk of becoming Long-term Unemployed (LTU), being able to intervene on their situation by assigning them the right resources.

## DATA OVERVIEW

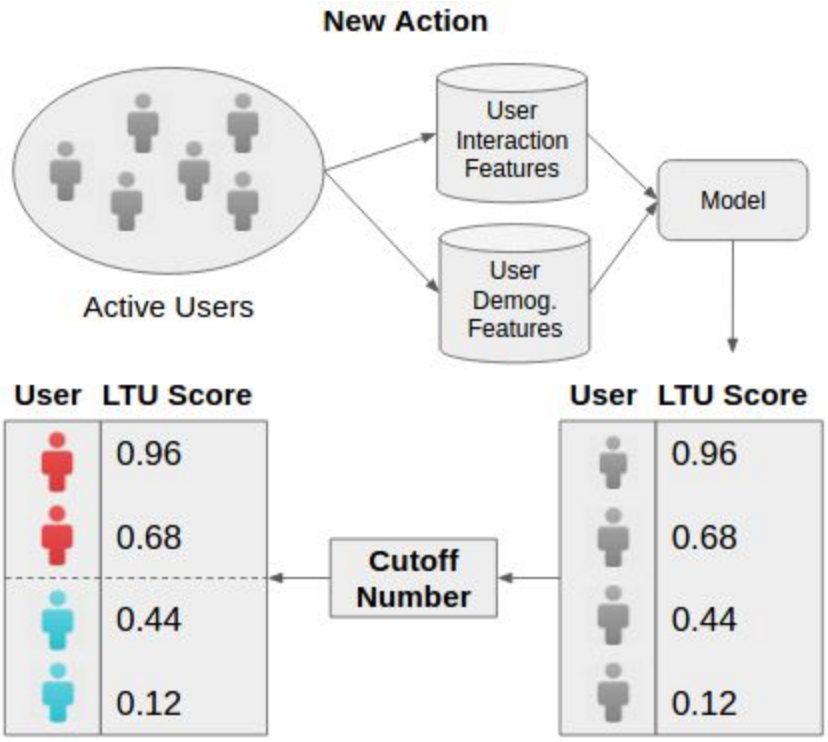
Portugal's Institute of Employment and Vocational Training – Instituto do Emprego e Formação Profissional (IEFP) provided 11 years of historical data about the unemployed population in the municipality of Cascais. The data contains the applications and interactions in IEFP unemployment system. Demographic information about the applicants and all their interaction with the IEFP system, such as trainings, job interviews, and meetings they attended were also provided. This adds up to a total amount of 125,000 applications, 74,000 unique users, and about 700,000 interactions with IEFP system.

## LTU RISK SCORE PREDICTION MODEL

Long-term unemployment is defined by Portugal's law as staying for at least 12 months active in the unemployment system. The individuals which fit this definition must receive special attention from the public institutions. In order to address this requirement, IEFP designed a model based on demographic information to predict the risk of becoming LTU when a new user registers in the system. However, this score is static, not being updated with time, and thus, it cannot be trusted after a certain time.

In this project, we propose a prediction model which can be run at anytime, also accounting for the interactions of the user with the system. This approach provides an updated LTU Risk Score, which can be used as additional information to prioritize users for different kind of interactions such as training or job placement.

Suppose the Municipality of Cascais wants to launch a  technical skills training for 50 people, and wants to prioritize the unemployed people at high risk of becoming LTU. They could run the model to assign an LTU Risk Score to the unemployed population and take this information in consideration in the selection proce.




## PRELIMINARY RESULTS

One of the outcomes of the exploratory analysis of the demographics was that applicants about 30 years old were much more likely to be LTU, as depicted in Figure 1.

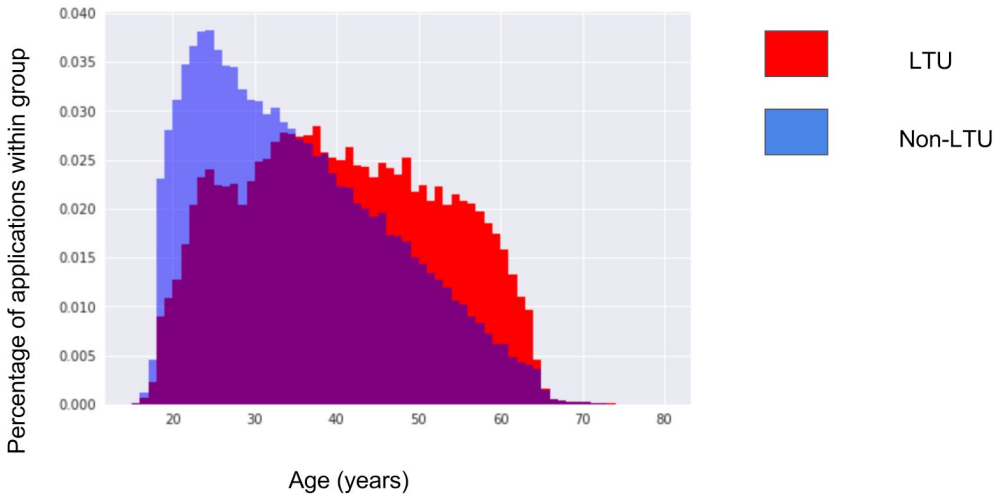



**Figure 1 -** Age Distribution between LTU and Non-LTU applicants

The age also came out as an important feature in all models tested so far. Using only demographic variables (without including any dynamic information), we trained a Logistic Regression and a Random Forest model, whose the performances can be seen in Table 1. In the test-sample, 29% of the applicants were LTU, which is the baseline for our model, and as can be seen, the model performed better than this.

| | Logistic regression | | | Random forest | | |
|---|---|---|---|---|---|---|
| Cutoff value | k=500 | k=1000 | k=2000 | k=500 | k=1000 | k=2000 |
| Precision | 52% | 48% | 43% | 43% | 42% | 40% |
| Important features | 1. Age<br>2. Number of dependents<br>3. Time since last cancellation | | | 1. Time since last cancellation<br>2. Age<br>3. Experience in previous profession | | |

**Table 1** - Preliminary Results for LTU Risk Score Prediction Model

## DATA SHARING BETWEEN PARTNERS

Another important goal of this project is help Cascais Municipality and IEFP in Cascais make better use of their data. Through this project, the Cascais Municipality, IEFP and other

organizations which work with unemployed people are sharing data and knowledge to be able to better assist the unemployed population on site. As we bring the partners into the discussions on how to build the solution and how it can be used, we can see a cultural change happening, which is also part of the goals of Data Science for the Social Good.

## NEXT STEPS

The next steps include:

- Add dynamic features as input for the model, such as: number of previous applications, outcome of previous interviews and training received in IEFP, etc.
- Apply more advanced feature engineering
- Run experiments to test different model types and parameters, feature sets, on different train/test windows
- Evaluate models performances using different metrics, such as: precision at K and ROC AUC
- Error and bias analysis
- Develop a prototype web application to show the results to partners