**SPRINGBOARD DATA SCIENCE CAREER TRACK**

**CAPSTONE PROJECT #2 – DRAFT**

**SILVIA MAIONE**

***Problem: Predicting Tanzania water pumps maintenance***

This project is based on a survey on water pumps conducted by the Tanzanian Ministry of Water. While typically the data has been used as an example of classification problem, here the main goal is to anticipate needed maintenance. Not only the probability of a water pump to be functional over the years can be predicted, but also preventive checks can be suggested. The stakeholders in this project are the government, any company or nonprofit agency involved in the management of the water pumps and the local population. Both the government of Tanzania and owners or installer of the water pumps can benefit from this kind of information, in order to perform predictive maintenance, which is not as costly as buying and installing new water pumps. The population also has great benefit since it can have access to water continuously. Other information can be shared with the stakeholders, such as the number of units and their conditions over the years and their geographical distribution. This allows to identify areas that need more attention.

The way the problem is posed makes it suitable for application of survival analyses and predictive maintenance concepts.

***Dataset description and data wrangling***

The dataset in this project has been initially provided by the Tanzanian Ministry of Water and Taarifa, an open source platform, and used to organize competitions having the goal of solving a multi-class classification problem. In fact, the main variable is the water pump status, which can be functional, nonfunctional or in need of repair. It can be downloaded from the following page:

<https://s3.amazonaws.com/drivendata/data/7/public/0bf8bc6e-30d0-4c50-956a-603fc693d966.csv>

The rest of the data contains the water pump attributes:

<https://s3.amazonaws.com/drivendata/data/7/public/4910797b-ee55-40a7-8668-10efd5c1b960.csv>

Each observation is a water pump, and there are about sixty thousand reported. The available features include geographical information of various type, such as region and basin, but also latitude and longitude, the amount of water and the population, and the year of construction and date of data entry. There are also many categorical variables, so, as part of the data wrangling process, these variables are encoded. Due to the high number of values, label encoding is chosen. Before even doing the encoding, missing values are dealt with. In fact, some variables which are believed to be relevant for the purpose of this project, such as the construction year, the amount of water and the population, have a high number of missing values. The time to event needed for survival analyses is created as a duration, defined by the difference of the recording and construction year. Since this is a needed information, and there are no elements to deduct the construction year from other features, the rows that do not have such entry are dropped. For the missing values in other features, other methods are used, for example taking the mean with respect to a certain area or the water pump status. Also, since the goal is to suggest maintenance before a water pump breaks, its status is encoded in only two levels, instead of three. At this point, any correlation between features is also verified, by using statistics such as Pearson and Cramers-V.

***Exploratory data analysis***

Once the dataset is cleaned up, it is ready for visualization. As shown in the following plot, most of the water pumps are functional, but when grouping those that are not with those in need of repair, the resulting two classes are balanced:

A screenshot of a cell phone

Description automatically generated

While the number of water pumps in a status group is similar across the various basins (areas with water resources), there’s more variation across the regions, and some of them have more nonfunctional water pumps, which indicates a need of maintenance.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Interestingly, while the number of water pumps has been increasing exponentially over the years, the number of those in need of maintenance has not changed much (this last information is based on the available knowledge of the construction year).

A close up of a map

Description automatically generated

This might suggest that there has not been a lot of maintenance over the years.

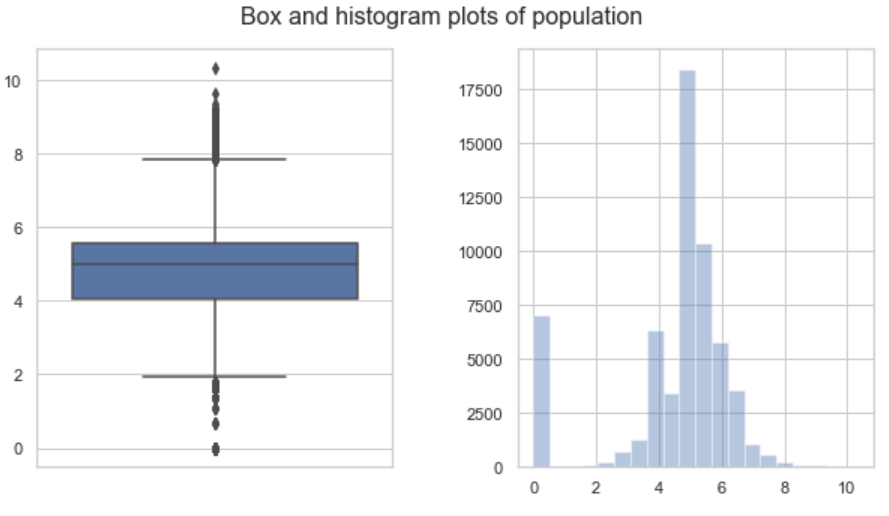
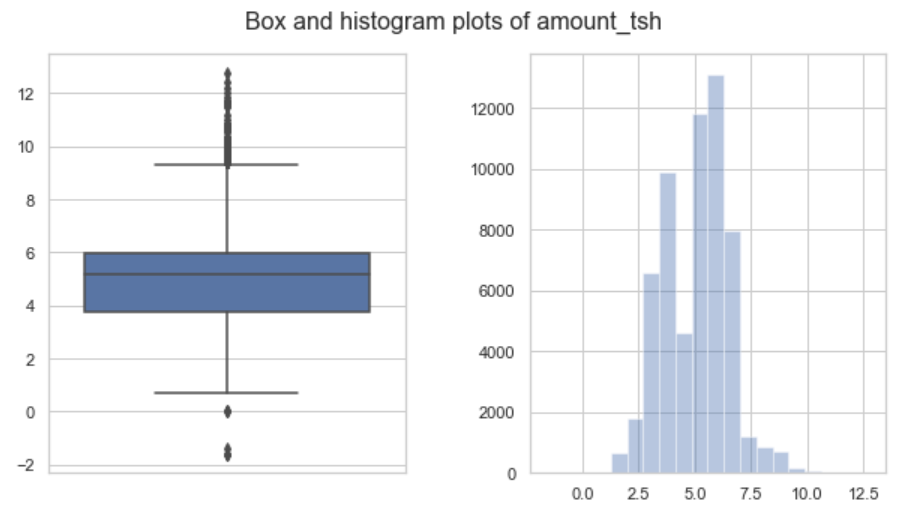
Some basins have not seen a significant increase of water pumps over the years. This can be an indication of lack of water resources or need of more investments.

A screenshot of a cell phone

Description automatically generated

The last insights can be shared with the interested authorities and agencies, not only to plan maintenance efforts but also possible future investments.

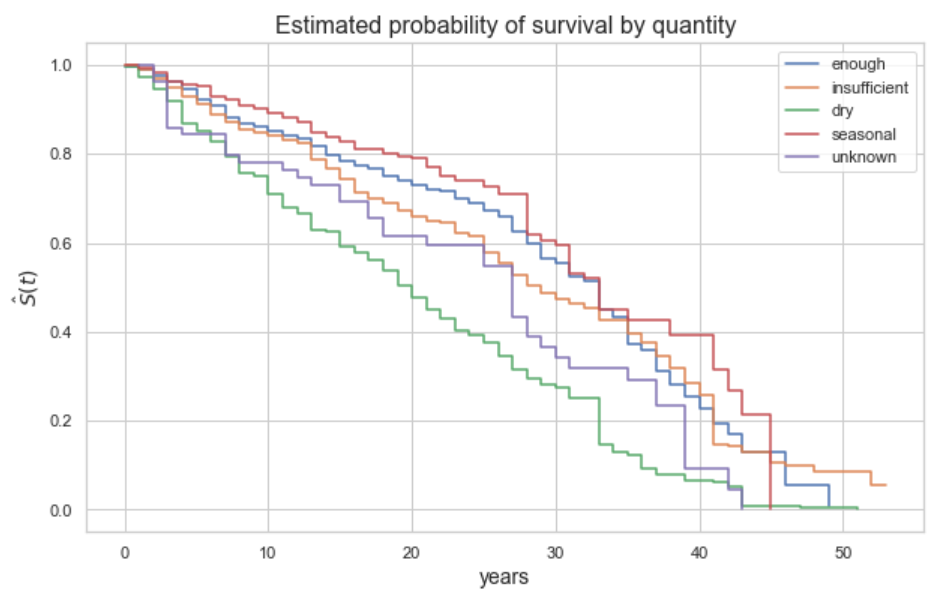
The numerical features such as population and amount of water are not normally distributed and are highly skewed. In the plots below they are shown in a logarithmic scale.



***Model***

For survival analyses a time-to-event is needed. The is defined as duration, i.e. the age of the pump at the time of recording. The event is the status, where a one indicates that the event occurred, so the pump is broken (or needs repair). The final dataset used to train the models has about one third of the initial available data, after the observations (i.e. water pumps) that have missing values for construction year, population and amount of water have been removed.

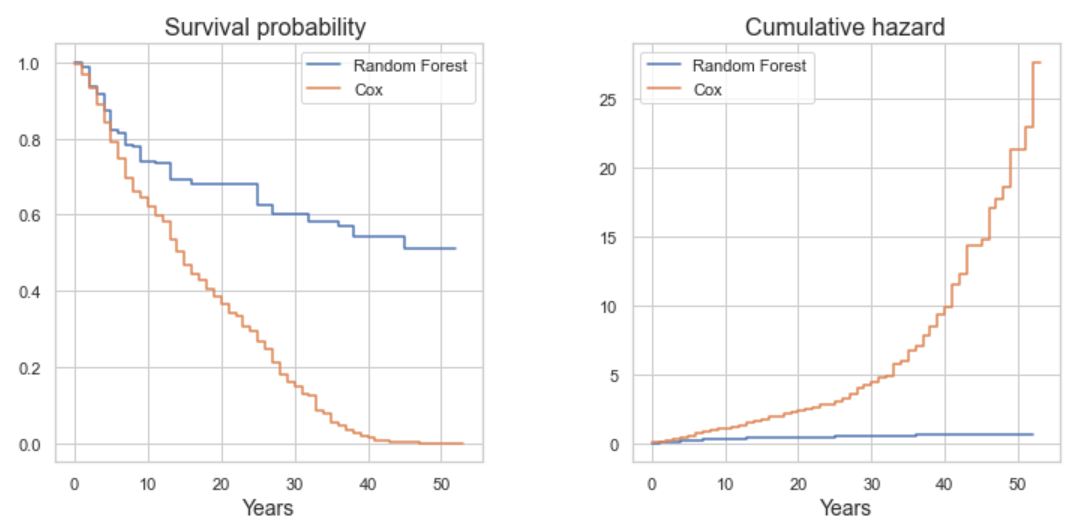
Initially, a simple Kaplan Meier model is trained. It shows the probability of survival, in this case of not breaking down, versus the time in years. So, it only uses the time and the event, but it can highlight differences when creating it with respect to different features, such as the water quantity.



In order to take into account all the features and build a more complex model, other possibility are explored, such as Cox regression, SVM and Random Survival Forest. Their Python implementation is available in the scikit-survival package, which is a module for survival analyses built on top of scikit-learn. The data is split in training and test set, the model fit, and hyper-parameters tuned for each algorithm. The metric used for evaluation is the C-index or concordance index, which is like the AUC of ROC in classification problems. In terms of performances, Cox regression and SVM with a linear kernel have similar performances, while Random Survival Forest is much better.

|  |  |  |  |
| --- | --- | --- | --- |
| **C- index** | | | |
|  | Cox | SMV | Random Forest |
| Training set | 0.629 | 0.583 | 0.867 |
| Test set | 0.624 | 0.582 | 0.797 |

Once the models are fit, a prediction can be made. For example, a water pump can be selected, and the survival probability and the hazard function can be plotted. The latter is an indication of the water pump to break down. Plotting such functions shows the difference between the models.



Instead of just removing all the missing entries, more data can be kept by filling the missing values of population and amount of water based on the geographical information of the non-null entries. If more data is used and the models fit again, Random Survival Forest results improve, while Cox’s do not change much.

***Conclusions***

In this project survival analyses concepts and algorithms have been used to suggest maintenance of water pumps in the state of Tanzania. The motivation comes from the economic and social benefits of having functional pumps and continuous water supply. The data used is a collection of observations made from 2002 to 2013 regarding water pumps built as early as 1960. While most of them are still functional, the number of those that needs repair has not changed much, possibly indicating a lack of maintenance. This is true in some regions more than others. The attributes used in the final models include geographical information, such as region and basin, and numerical, such as population and amount of water, plus others like quantity and source of water. Among the three algorithms tested, Random Forest applied to survival analyses proved to be the best performer, while Cox regression is not able to learn the complexity of the data.

***References***

1. <https://github.com/aspds18/Springboard_capstone2>
2. <https://scikit-survival.readthedocs.io/en/latest/api.html>
3. Pölsterl, S., Navab, N., and Katouzian, A., [Fast Training of Support Vector Machines for Survival Analysis](http://link.springer.com/chapter/10.1007/978-3-319-23525-7_15). Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, Lecture Notes in Computer Science, vol. 9285, pp. 243-259 (2015)
4. Pölsterl, S., Navab, N., and Katouzian, A., [An Efficient Training Algorithm for Kernel Survival Support Vector Machines](https://arxiv.org/abs/1611.07054). 4th Workshop on Machine Learning in Life Sciences, 23 September 2016, Riva del Garda, Italy
5. Pölsterl, S., Gupta, P., Wang, L., Conjeti, S., Katouzian, A., and Navab, N., [Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients](http://doi.org/10.12688/f1000research.8231.1). F1000Research, vol. 5, no. 2676 (2016).