MACHINE LEARNING – Assignment 2
M.SC. IN DATA SCIENCES AND BUSINESS ANALYTICS M.SC. IN ARTIFICIAL INTELLIGENCE
CENTRALESUPELEC

Flora Attyasse Vadim Benichou David Kiskovski Marc Terzikhan

Shipwreck of Titanic

On the 15th April 1912, the liner boat Titanic sank after colliding with an iceberg, and the "Women and children first" rule regarding survivors' evacuation was observed. To go beyond, an attempt will be made at using Machine Learning techniques applied to the dataset of passengers and their characteristics to predict specifically what made them more likely to survive.

The raw dataset comes from a Kaggle competition (*Titanic: Machine Learning from Disaster*) and is composed of a train and test set respectively of size 891 x 12 and 418 x 12, each row being a passenger.

Exploratory Analysis – Feature Analysis – Feature Engineering

The first audit of the datasets started by looking for outliers and missing values. Outliers were found on the train set using the Tukey Box-Plot method. An outlier is an instance with at least two outlying values for the numerical features 'Age', 'SibSp', 'Parch' and 'Fare'. Another method could have been a clustering with deviation degree or a distance based few neighbors.
Seven passengers have very high value of SibSp and the three others a high ticket Fare. Since least squares estimates for regression models are highly none robust against outliers, the above instances were dropped as they might strongly impact the predictions.

The train and test sets were then merged to maintain consistency during the preprocess, consisting of dealing with missing values and switching categorical encoding into numerical values.

Table 1: Missing values in whole dataset

Age	256
Cabin	1007
Embarked	2
Fare	1
Survived	418

The train set was analyzed by plotting features (below) or a combination of features seeming significant against the target variable 'Survived'. This was performed in order to get a better grasp over the features in order to later create even more significant ones.

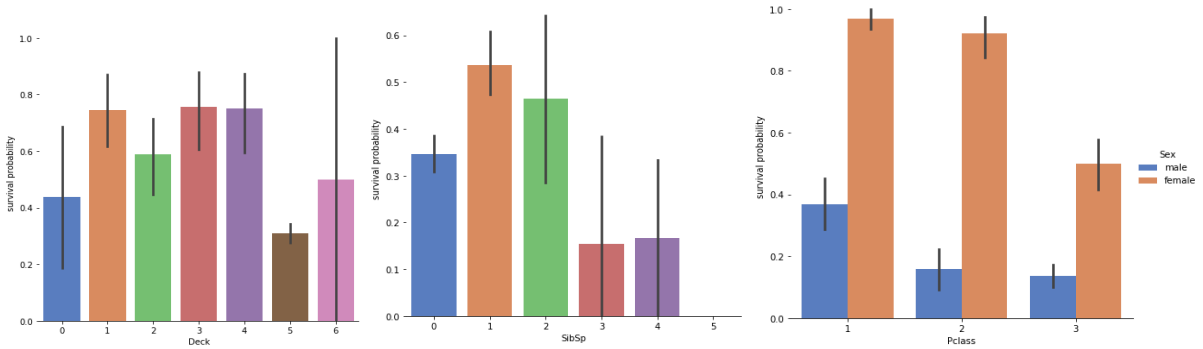

Figure 1: Plots of probability of survived vs various features

Plot Analysis

It was found that age is a factor of whether a passenger survived or not. Young passengers seem to have managed to survive more as opposed to passengers between 60-80 years old. As expected, females also had a much higher probability of survival, and so did people in higher classes.
The deck plot, simple extraction from 'Cabin', doesn't show any clear relation to survival.
Passengers having a lot of siblings/spouses also had less chances to survive and single passengers (0 SibSP) or with two other people (SibSP 1 or 2) have more chance to survive.

Engineering

A feature called 'Title' was created by extracting the title of passengers from 'Name'. This was done in order to add an additional characteristic to the passengers which could symbolize their "importance" in society, hence maybe contributing to their survival chances. However, the main reason this was performed was to deal with the missing values of the "Age" feature. Indeed, a passenger who's age was absent from the dataset was attributed an age equal to the mean of the category of "Title" he or she belonged to. This makes sense because titles are also highly correlated with "Age". Some passengers had rare titles, such as "Dona", "The Countess", "Dr", etc… Because of the very limited amount of these titles, which seemed like an indication of an elitist social class, they were initially all grouped in a title called "other" as it would make sense that if wealth and social ranking are factors for survival, then they would have a very high chance of survival. However, in the end each special title was added to the most common title for its gender group ("Dona" for Mrs and "Col" for Mr for example). There is no ambiguity as to whether some titles like "Dr" would correspond to a male or female, as females at the time did not have the same right to bear some professions as their male counterparts. This modification was made as to avoid any overfitting that could be done due to the very limited amount of "other" titles that are present. This modification ended up improving the accuracy of the overall model.
The 'AgeRange' feature was also added by binning the ages into four categories: Toddler, Child, Adult and Elderly. This was done in order to properly take into account how people of different ages had various chances of survival in the final model.
'Embarked' missing values were filled with the attribute mode. As the most frequent location is Southampton, the missing values were filled with 'S'. Both missing values were also found to be the same ticket ID, hence the assumption that both passengers embarked from the same port seems very reasonable.
Now the instance having only missing value of 'Fare' belongs to Pclass 3. It was hence assigned the median ticket fare of the class.
A family size 'Family' feature was created by adding 'SibSp', 'Parch' and the passenger himself. This was done as it isn't a stretch to think that the more people someone had to care for, the less his chances of survival were, as previously discussed above.
A new feature 'FareC' was also added as to group the amount of money paid for a ticket into multiple categories ranging from below average, average, over average and expensive. These intervals were chosen by inspecting the distribution of the 'Fare'. It could differentiate the survival rate of people from the same class.
Finally, a 'Deck' feature was created by extracting the first letter of the 'Cabin' feature. This signifies where in the ship the Cabin was. This was thought to be a good idea as it is easy to see how a portion of the could have started to sink first hence people from that place would have a higher probability of death. To deal with the (more than 70%) missing values associated with this column (from 'Cabin'), a decision tree was performed with respect to other relevant features such as 'AgeRange', 'FareC', 'Title', 'Pclass', 'Family', 'FareC', 'Embarked' (not 'survived' as this is what we want to find). This feature ended up having very limited to no impact on the model. If more time had been allowed, it would have been interesting to look at the numbers of individual cabins and try to work with them, as instead of a 'deck' which symbolized the level on which the cabins rest, the number could indicate on which side of the ship they are. It would then be expected that since the ship started sinking on one of the sides, the survival rate of people staying on that side would be lower on average.
The features: 'PassengerID', 'Name', 'Age', 'SibSp', 'Parch' 'Cabin', 'Fare' were then dropped as they were either redundant features to the ones we had created or simply could not be used in any practical way (such as Name).

Table 2: Discretization and binning

Feature	Transformation
‘Embarked’	'C': 0, 'S': 1, 'Q':2
‘AgeRange’	‘Toddler': 0, 'Child': 1, 'Adult': 2, 'Elderly': 3
‘Title’	'Mr': 0, 'Miss': 1, 'Mrs': 2, 'Master': 3
‘Sex’	‘male’: 0, ‘female’: 1
‘FareC’	‘0' if ‘Fare’ <35’, '1' if in [35,75), '2' if in [75,110), ‘3’ else
‘Deck’	‘A’:0, ‘B’:1, ‘C’:2, ‘D’:3, ‘E’:4, ‘F’:5, ‘G’:6
‘Family’	‘SibSp’ + ‘Parch’ + 1

The use of integers in the categorization of the data was to enable algorithms such as decision trees to be performed on the data.
Following this, the dataset was clean and could be divided into the original train and test datasets again.
Finally, as all features were now numerical, the correlation heatmap could be computed.

Modeling

Multiple classifiers were used and compared for the model their associated accuracies were computed by a validation method on a validation set obtained by splitting the train set. Three splits were performed: 10%-20%-30% for the validation.
The models considered were Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Decision Tree, Random Forest and Gradient Boosting.
Ridge Regression and LDA were performed but finally excluded from the final model as they did not raise its accuracy, but the methods were nevertheless useful to estimate the inference of each feature upon the target variable.

Table 3: Models Accuracy on validation dataset

Model	Accuracy
Gradient Boosting	88.14
SVM	86.44
Decision Tree	86.44
KNN	84.75
Regression	84.75
Random Forest	81.92

All methods were also combined into a StackingClassifier ensemble method but the resulting accuracy was smaller than each model on its own, this was perhaps due to the fact the other methods were not “weak” enough, which is required for these types of models.
All the models were tested directly on the test dataset from Kaggle because each model displayed satisfying accuracy on the validation dataset. The one chosen was the Support Vector Machine with an accuracy of 0.79425 on the test dataset (Prediction_Titanic_30SVM.csv).

Conclusion

According to our predictions it seems that the survival is most associated with the age, sex, family size and the social rank of the passengers. Indeed, upper-class passengers were more likely to survive. This machine learning project centered on the isolated incident which is the sinking of the titanic goes to show that various factors impact the very primitive question of whether you will live or die. If any one of the men who died had been born a woman instead, they probably would have survived. But survival wasn’t entirely based on such fatalistic reasons. Like many over things in life, people living in wealth were also amongst the most privileged and more were allowed to live as opposed to other people with modest backgrounds. Although machine learning projects such as this one often turns the death of innocent people into simple numbers of 1s and 0s, they can however be used to answer some important as to “what is it that made someone survive and someone die ?”. The socio-economic implications of these answers can in turn be studied in order to find out more about ourselves as a species.