Udacity - Machine Learning Engineer Nanodegree Program

Capstone Project

Matheus Sena Vasconcelos
Fevereiro 28st, 2020

I. Definition

I.I. Project Overview

RMS Titanic was designed to be the more luxurious and safest ship built in 20th century. On the night of April 20, the Titanic hit an iceberg and sink in the middle on its journey. Unfortunately, due to the low number of rescue boats, more than a half of the passengers have died. The survive number was only 722 of 2224 in total.

The project proposal is to build a predictor model that recieves, as input, passenger information (like name, age, gender, socio-economic and class), makes the text preprocessor, guesses if this fictitious passenger would survive or not in Titanic tragedy and return it as a HTTP response. As an experiment, supervised and unsupervised machine learning algorithms will be used to build and improve the model. To go further and receive theses passenger information, an endpoint will be develop using Python Frameworks in order to demonstrate another away to create endpoints, instead of those shown during the Nanodegree Program using AWS.

The main idea of this project is to put into practice all the machine learning and software engineer knowledge learned during the Machine Learning Engineer Nanodegree Program and join it into my developed skills as a Software Developer.

I.II. Problem Statement

The problem is Kaggle challenge and can be access [here (https://www.kaggle.com/c/titanic)](https://www.kaggle.com/c/titanic). Based on the passenger data, the challenge is to build a predictive model that answers the question:

> “what sorts of people were more likely to survive?”

In others words. The idea is to use the provided dataset, which contains all informations about the passenger aboard Titanic in 1912, and build a machine learning model that predicts if the passenger would survive, based on new data received.

To build a great predictor model, supervised machine learning classification algorithms will be used, like K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM). Due to the labeled dataset provided, which indicates if the passenger survived or not.

I.III. Metrics

For the model evaluation, the follow metrics are used to measure how good the model is.

- **Accuracy Score**: value that indicates how many predicts the model guessed that the passenger would survived and guessed right, comparing to the total data sent. In other words, is the total number of True Positive and True Negatives divided by the total number of samples.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}}$$

- **Confusion Matrix**: matrix that indicates how many True Negatives, False Positives, False Negatives and True Positives. Our goal is to increase the number of True Negatives and True Positives, which show that the model more guessing right than wrong.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

II. Analysis

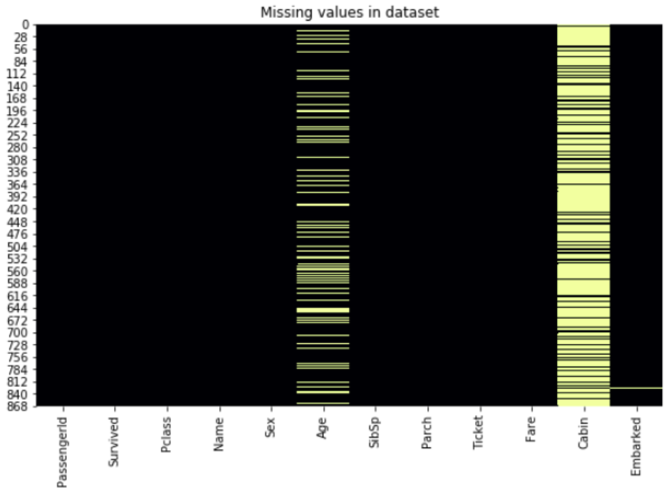
II.I Data Explation

II.I.I Dataset

The dataset brings informations about passenger onboard on RMS Titanic that have survived or not on the night of tragedy. Each row contains unique passengers with different information about them, from name and parents onboard to the amount paid on the ticket.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Over the 890 rows in the dataset, 866 cells have missing values and the most of those values are in the Cabin column. Due to the quantity of null values in that column, it can't be used as features to the model. In other hand, features like Age, which also have null values, but they can be filled out with mean or median. The follow heatmap plot shows that in more details.



Whites lines indicates how many values are missing in each column. It's clear that Cabin column has almost all of its values as empty. While Age column has less than half and Embarked column only one or two data.

II.I.II Data Statistics

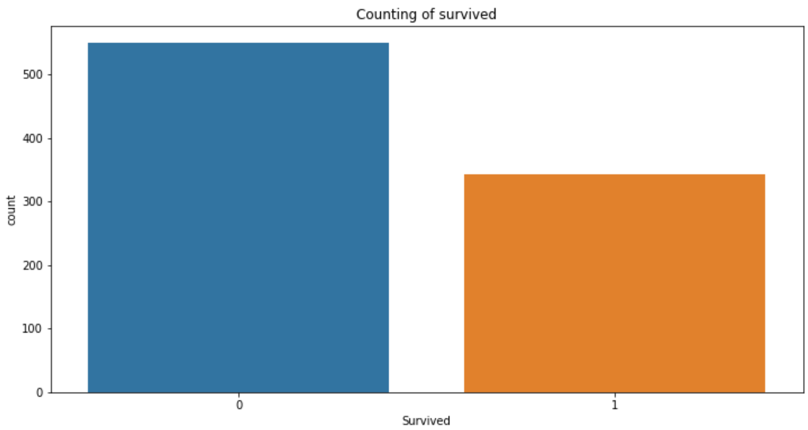
With Pandas, a Python Library, we can easy extract statistics information in numerical columns in a dataset.

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

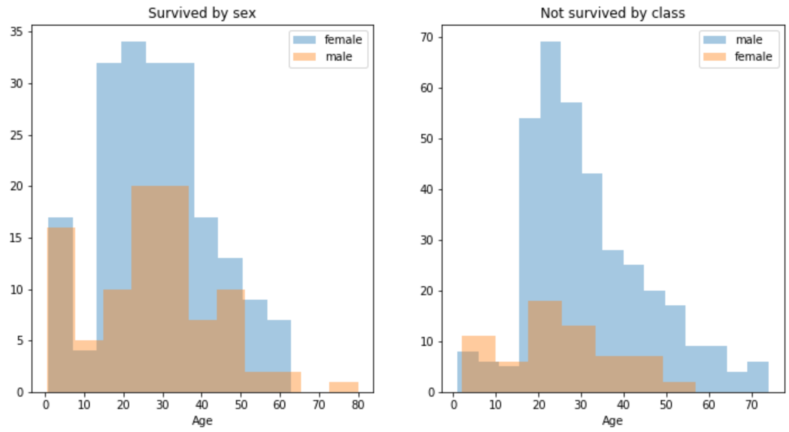
We can see useful information in the table above. It's noticed that the mean of all ages are equal to 29.69 and the missing values in Age column can be filled with it. Another interesting information is the Survived mean value, only 38.38% of the passenger have survived. It's indicates that we have an imbalanced dataset and have to balance it in data preprocessing step.

II.II Exploratory Visualization

It was said before that the dataset is imbalanced and there are more not survived passenger. The next plot show this counting the number survived and not survived passengers. 0 indicates not survived and 1 indicates survived on the X axis.

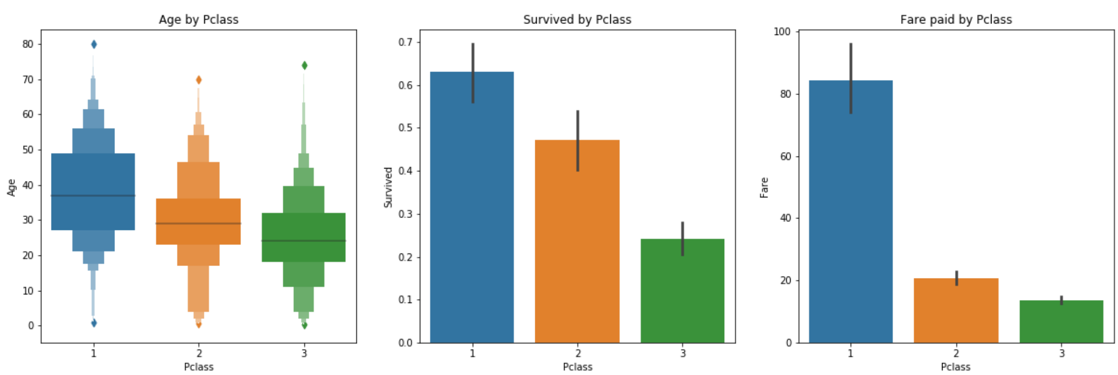


We can go further and see how many person by sex have survived based on their age.



The next plot brings a lot of information about the dataset. The first figure (Age by Pclass) shows the mean, median, min, max and quarters values of the age by passenger class. This kind of plot is useful to use to fill missing values in the Age column. To not drop all null data in Age column, we can fill it with mean age per class, without spoil the entire dataset.

Other information we can extract of figure is which class had more survived passenger. The Survived by Pclass plot shows that the class with the greter number of death was the Pclass 3. This kind of information indicates that, when the rescue boats arrived after the crashed, they prioritize the more fancy class (Pclass number 1). To reiterate this information, the Fare Paid by Pclass plot show that Pclass 1 had the most expensive fare and, consequently, Pclass 3 the cheapest.



II.III Algorithms and Techniques