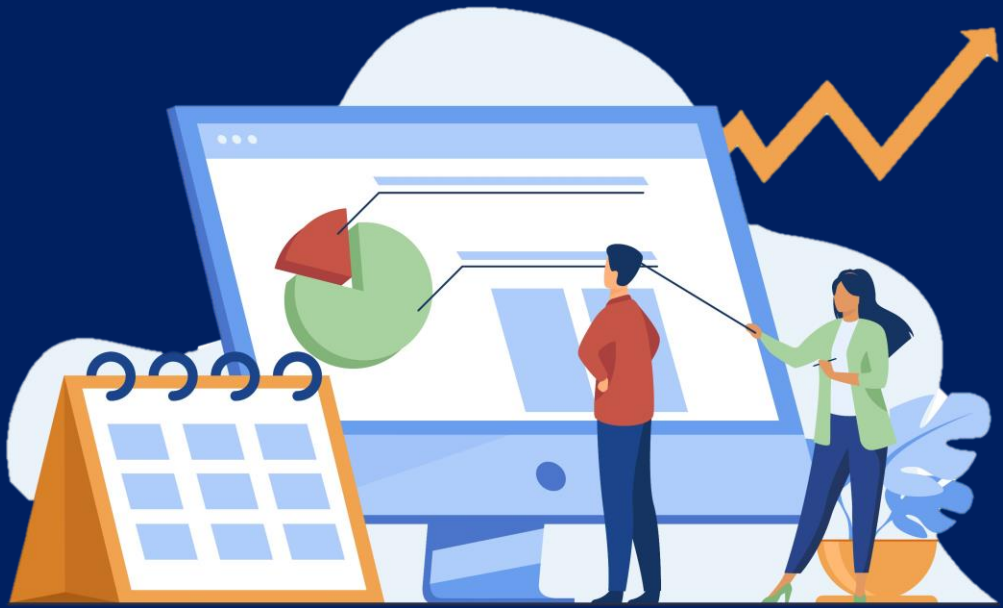
Let's Analyze

Titanic : Who do you think survived?



Introduction

We attempt to analyze the classic Machine Learning Problem – Survival Prediction on the Titanic, in a simpler way, for anyone who is getting introduced to analytics.
The dataset is available on public domains, for anyone to try their hands on.
Let's see how the numbers work out – Let's Analyze!

Process

- Import the data into Excel / Python, and explore the available variables, missing values and preliminary insights
- Analyze the data to identify patterns, variables that have an impact, and the magnitude of impact
- Fit a model to allow prediction using algorithms and parameters – It could be anything from linear modelling to supervised machine learning

Exploring the Data

The data is available on public platforms like www.kaggle.com, fortunately in a csv format – helpful for most of us, isn't it? Download and import the same in your software of choice – Excel / Python / R / Others. You will notice 3 files – train, test and sample submission. We can ignore the last 2 and focus on the train data for now. The table should look like this:

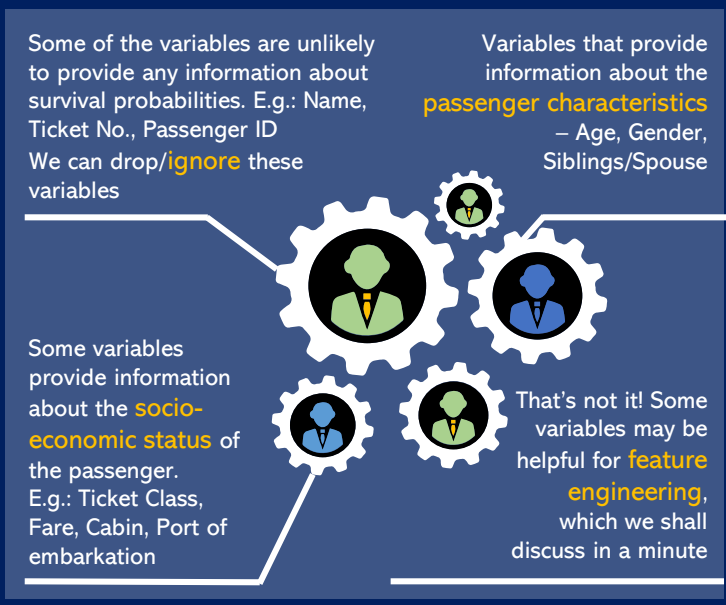
PassengerID	Survived	Pclass	Name…
1	0	3	Braund…
2	1	1	Cumings…
3	1	3	Heikkener…

… and up to 12 variables. Let's see what each of them mean :

- PassengerID : A unique identifier for each passenger
- Survived : Did he/she survive? Yes = 1, No = 0
- Pclass : Ticket class,1 being the most expensive
- Name : Passenger Name
- Sex : Passenger Sex
- Age : Age of the passenger
- SibSp : # of siblings/spouses on board
- Parch : # Parents/children on board
- Ticket : Ticket Number
- Fare : Price of the passenger's ticket
- Cabin : If they had a cabin, the cabin number
- Embarked : Port of embarkation on board
    - S – Southampton
    - Q – Queenstown
    - C – Cherbourg

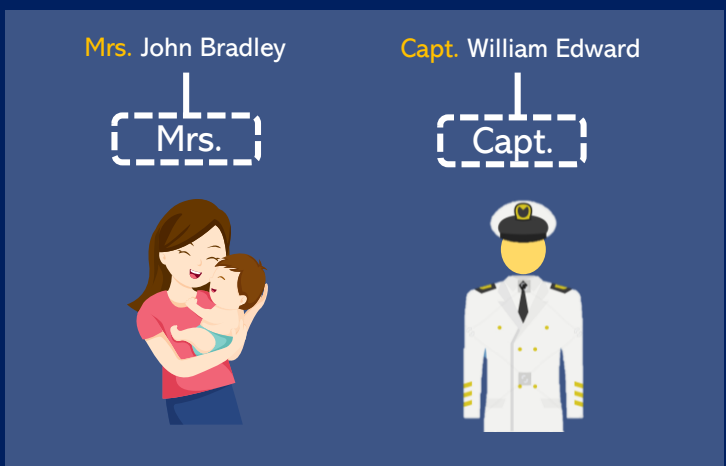
What information does the data provide?

The "Survived" variable is what we aim to predict. All other variables may provide information that leads to an accurate prediction – these are called "features" (In statistics, Independent Variables). Categorizing the variables into informative buckets may help us extract valuable information



Feature Engineering

Sometimes, a variable may seem like it cannot provide information, but anyway might have a lot to offer. Modifying/Engineering a variable to extract information is known as "Feature Engineering". Consider the following example –
Did you think "Name" cannot provide information? Consider the following 2 names, and try this :



Notice the titles :

- "Mrs." tells us about a married woman, who is likely to have her husband on board to protect her
- "Capt." is likely to be a captain of the ship, who, unfortunately, is likely to sacrifice his life trying to keep the ship afloat.

Similarly, we extract the titles of all names in the data and call it a variable "Title". Hopefully, it shall help in our analysis.

What else can we engineer?

Siblings/Spouse + Parents/Children + 1 =Family Size

Note: we add 1 for the passenger themselves

Call this variable "Family Size"

If Family Size = 0,

Then Passenger is "Alone", else not

We create a variable "Alone", where :

Alone = 1 , if Family Size = 0

Alone = 0 , if Family Size > 0

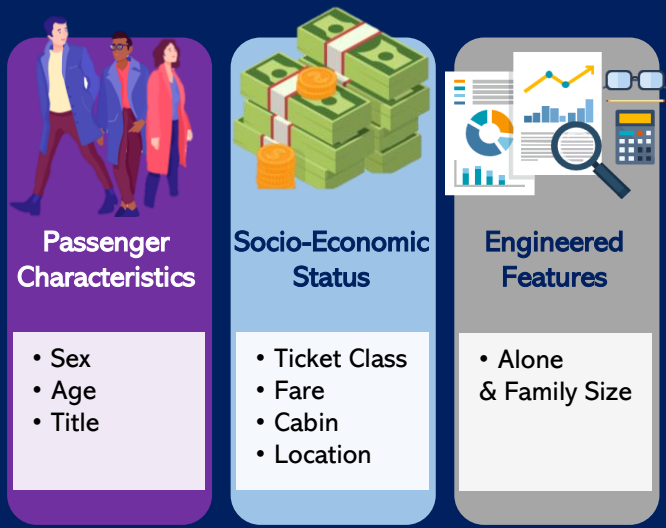
Missing Data

For missing values in data, like you may find in "Age", take a smart estimate for all data points that are missing. For example, you could assume the average/median of all other ages as the estimated value of missing age data.

Note : People acquainted with data science could use classifiers, regressors or other imputation algorithms. Here, I've used a Random Forest Regressor to estimate the missing ages (Just a way to guess a close number!)

Analysis

Now that we understand our data, and we have modified it to extract enough information, the next step is to proceed with the analysis of this information. Previously, we'd categorized the variables into buckets that gave us relevant information about the survival probability of a passenger. Let's analyze each of these categories and start visualizing information.

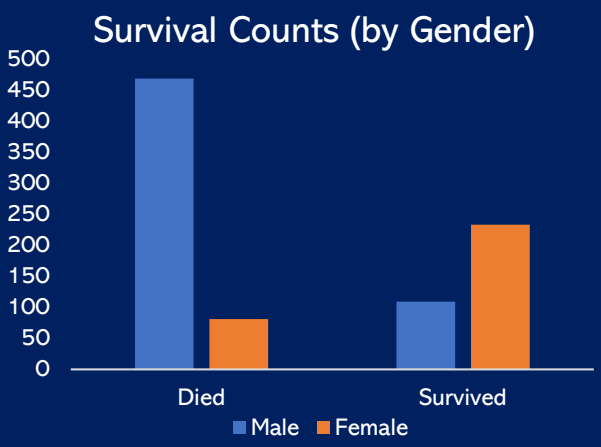


Passenger Characteristics

Passenger characteristics define natural attributes of any passenger, which may/may not have had an impact on the survival of the passenger. However, external factors and circumstances may lead to strong patterns arising among passengers with different characteristics.

Characteristic 1 – Gender

We plot the number of people that survived/died in the wreckage, split by gender.



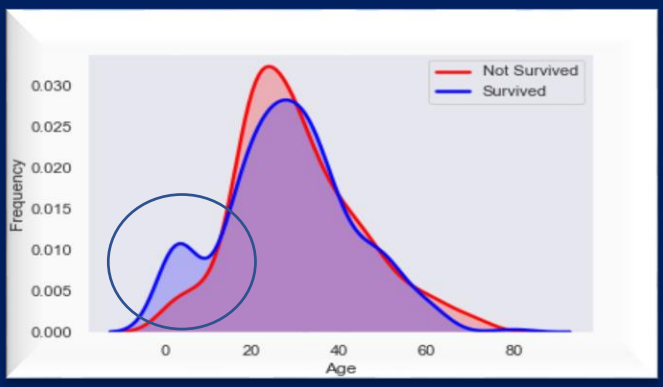
Insight – there was a higher likelihood of a woman to survive than that of a man.

Working the numbers, we find that only **19%** of the men survived, whereas **74%** women lived to tell the tale.

Characteristic 2 – Age

We plot the frequency (# of people of that age/Total Passengers) of people that survived/died in the wreckage, across all ages.

Note : For those using python, a seaborn kde-plot can help visualize such curves for continuous values i.e. Age



Notice the bump for survived passengers of ages 0-10

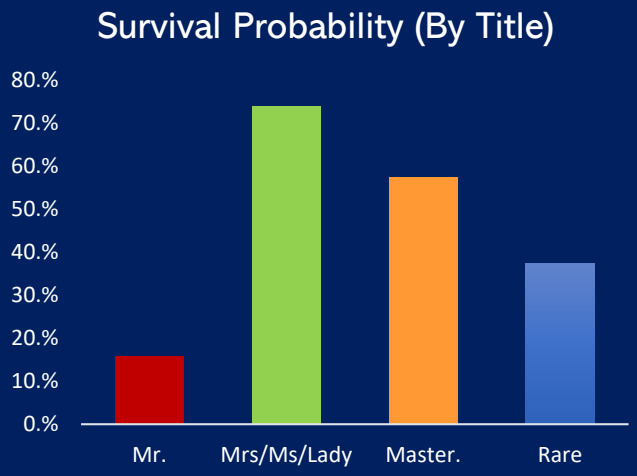
Insight – Infants had a higher probability of surviving, as compared to all other ages.

Note : Since we're plotting frequency here, the rise/fall of the curve also accounts for the proportion of people of the corresponding age in the entire list of passengers. Hence, ideally, we would expect both the curves to rise/fall along with each other. It's the deviation at ages 0-10 that helps us notice a pattern.

Characteristic 3 – Title

We plot the probability (% of people survived) based on the 'Title" variable that we derived from the names.

Note : We have clubbed ["Mrs.", "Ms.", "Lady"] into one group for simplicity. Titles like ["Capt.", "Col.", "Dr.", "Rev.", "Don.", etc.] have been categorized as 'Rare'



Insight – As we've seen before, women (Mrs./Ms./Lady) and children (Master.) had a higher likelihood of survival, whereas adult men (Mr.) were more likely to die. It is interesting to notice that "Rare" titles had an adequately high probability of survival – a possible explanation for this is that ["Dr.", "Col.". "Don.", "Rev.",etc.] are usually doctors, colonels and other socio-economically powerful people, who might have enjoyed privileges.

Key Inference –

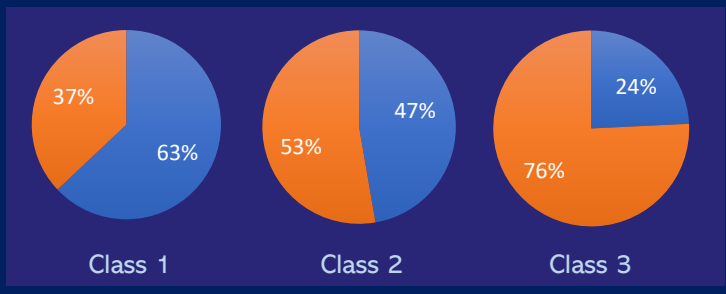
Women & Children were highly likely to survive the wreck.

Socio-Economic Status

We have gathered enough information about how passenger characteristics may have helped their case. However, it is quite likely that some passengers, because of their influence, enjoyed privileges that others did not. Some were prioritized, some were not. Let's consider the variables that act as indicators of socio-economic status.

Characteristic 1 – Ticket Class

We have information about what class of ticket did a passenger hold – 1,2 or 3. Tickets to board the Titanic were sold not only to the rich but also to the other economic classes. However, the passengers were tagged among one of the 3 ticket classes. Furthermore, the classification was based on their wealth as well as social status. Hence, this is likely to be a good socio-economic indicator. Let's visualize the impact this had on survival probabilities

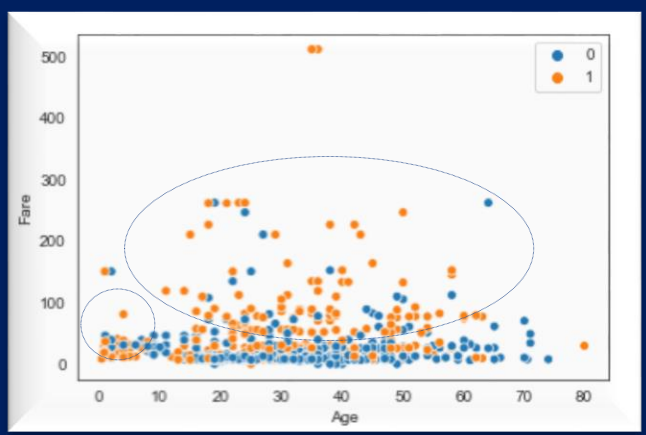


Notice how the survival probabilities are falling through each of the ticket classes.

Insight – Not to our surprise, the passengers with Class 1 tickets were most likely to survive (63%), whereas the passengers with Class 3 tickets were least likely to(24%).
In all probability, the rich and influential people were prioritized when it came to lifeboats and life-jackets.

Characteristic 2 – Fare

While plotting fares, we need to be careful that fares aren't related to other variables (In Statistics, Multicollinearity). For example, if kids were charged higher fares, our data would show that higher fares led to survival, which wouldn't necessarily be true.
Hence, we create a plot of Fare vs. Age, and examine the data-points that survived.
Note : Python users can use the seaborn scatterplots with hue for this exercise.

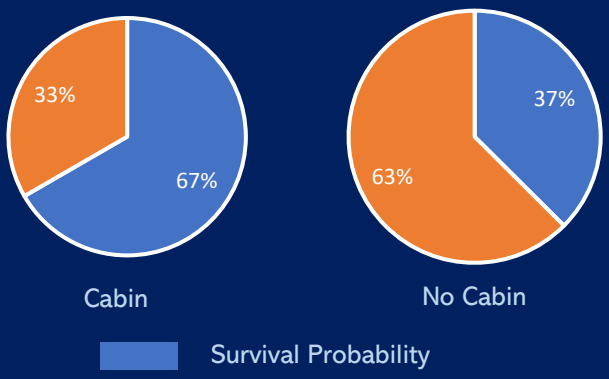


Notice how there is a surge of orange dots (survived) towards the upward section of the plot, and that of blue dots (dead) towers the lower section. Also, there is a surge of orange dots (survived) toward the origin, where age is 0-10, reconfirming our previous analysis.
Note : we can confirm that fares were not related to age, since we see no horizontal pattern between Fare vs. Age

Insight – Passengers who paid higher fares were more likely to survive than those who paid lower fares.
We also notice 2 middle-aged outliers, who paid 500 units in Fare, and both survived. They're likely to have been people from the royal circles.

Characteristic 3 – Cabin

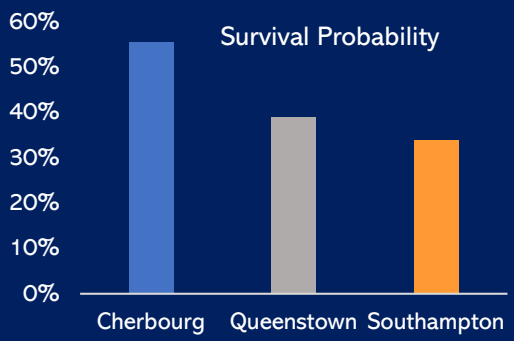
We plot the survival probabilities of people with and without cabins in pie charts.



Insight – We observe that passengers with cabins had a higher likelihood to survive the wreck.
Further analyzing, we notice that the average fare of a passenger with a cabin was 76 Units, whereas for the ones without a cabin, it was just 19 Units

Characteristic 4 – Location

We plot the survival probabilities of people who'd boarded the ship from different locations.



Insight – We observe that passengers from Cherbourg had a higher probability to survive, over 50%.
A possible explanation for this can be the economic prosperity in Cherbourg at the time, and hence wealthier individuals. However, this must be further investigated using data.

Reconsider - who do you think would have lived **Jack or Rose** ?

Engineered Features

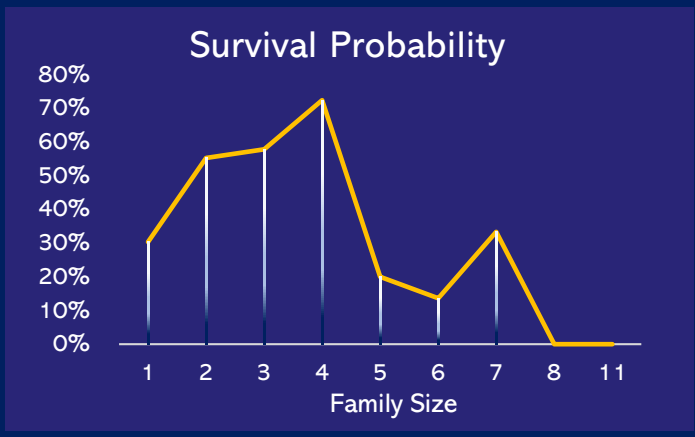
Characteristic 1 - Alone & Family Size

We try to work out some statistics on survival probabilities of people who were alone / with family. If they had family on board, we might get further information by analyzing the size of the family.

Not to our surprise, we observe that the survival probability of passengers travelling alone was just 30%, as opposed to 51% for those who were not travelling alone.

One can rightfully expect that having family on board increases the likelihood of survival – help matters! However, what if a passenger had too many people to help? Hence, we hypothesize that although having a small family helps, having a larger-than-average family might be detrimental to survival during a shipwreck.

To validate this hypothesis, let's plot the survival probabilities by family size.



Insight – The pattern in the curve strengthens our hypothesis. Having a small family helps you survive. However, someone trying to protect a large family might not be able to protect themselves.

Modelling

Now that we have gathered all the information we needed; we can proceed with fitting a model to all the data that we have.

1) **Prepare the data** : Convert all categorical variables to numerical data, for any model to understand. (E.g.: Male = 1, Female = 0). For those acquainted with statistics, dummy variables are a good way to quantify categorical variables.

2) **Fit any model** of the following :

- Simple Linear Regression, that calculates coefficients for each variable's impact on survival.
- Logistic regression, that uses the logistic function to model binary (Yes or No) variables.
- Decision Tree/ Random Forest Classifiers, that create decision trees containing each variable at various nodes.
- Any other model that seems fit

3) **Test the model** by running it on sample test data, check the accuracy by picking data points at random to check whether the prediction matches the reality/expectation. An accuracy score can be measured by calculating the proportion of correct predictions.

In most cases, no one model is the perfect solution. One can use whichever model they understand best and find appropriate. For more information on any of these models, adequate material is available publicly. I've used a Random Forest Classifier here, that gave me an accuracy of over 85%.

Wrapping Up

The dataset is available at https://www.kaggle.com/c/titanic/data

For analysis, use any software of your choice. Here, I've used a combination of Python & MS-Excel, since plots from Excel are much more flexible and compatible with a presentation. However, plots from python as well can be modified and used for this purpose. Similarly, Data analysis could be conducted on any platform.

For more knowledge on any of the terminology or models, adequate data is available on public domains, which should be easily accessible from any search engine (E.g.: Google)

Trying this hands-on can prove to be a helpful exercise if never done before.

For any further queries/information, please reach out at

Vaibhav Agarwal, FRM
vagarwal4765@gmail.com

LinkedIn : https://www.linkedin.com/in/vaibhav-agarwal-frm®-207390ba