Business Understanding

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

The challenge here is - Given a passenger's information, how can we predict whether he/she survived the Titanic disaster?

Analytical Approach

Our target variable is categorical (survived / not survived), and hence we need classification models for this task.

Data requirements

We would require onboard passengers information which might include name, age, fare, gender, class.

Data collection

We are given two datasets both of which are CSV files, one for training our model named as train.csv and the other test.csv to test if our model can determine survival based on observations, not having the survival info.

Data Understanding

This step is part of Exploratory Data Analysis

There are 891 observations in the training dataset with each having 12 columns. 11 of them are predictor variables and 1 being target variable.

There are few different types of variables available.

- •Continous: Age, Fare
- •Discrete: SibSp, Parch
- •Categorical: Survived, Sex, and Embarked
- •Ordinal: Pclass
- •Mixed: Ticket
- •Alphanumeric: Cabin

There were 3 features having missing values.

- •Cabin
- •Age
- •Embarked

Cabin has way too many missing values and hence it is better to drop.

As per the training dataset, there were more male present compared to female and most of the people didnot survive. But females had better survival rate than males. It was also found out that survived passengers had paid more fare than the ones that did not survive. At the same time Pclass = 1 had better survival rate than the rest 2 classes.

Majority of the passengers were between the age group 15-35 but most of them didnot survived. Children aged < 4 and old aged people had higher survival rate.

Data Preparation / Feature Engineering

After closely looking into the dataset, variables types, values, amount of missing values present, I have decided to

- Impute the missing Age values
- Turn age into an ordinal feature
- Impute the missing Embarked values
- Drop Cabin [too many missing values]
- Drop Ticket [many duplicates]
- Drop PassengerID, Name, SibSp, Parch [not helpful]

I also performed some feature engineering as there are few categorical variables present. Created Dummy Variables for

- Sex
- Embarked

Modeling

Models trained

- Logistic Regression
- k-Nearest Neighbors
- Support Vector Machines
- Naive Bayes classifier
- Decision Tree
- Random Forest

Evaluation

Decision Tree and Random Forest achieved the maximum accuracy of 93.03%. Hence, I recommend using them for this task.

Further Improvements

I can use cross validation and Sampling techniques to improve the accuracy. Also I can work on creating more features rather than dropping them which might increase the accuracy.