Titanic Survival Prediction: Exploring Data to Save Lives!

Skill- DSML

Introduction

Data Collection and Preparation

EDA

Feature Engineering

Modelling approach

Results and Evaluation

Conclusion

Introduction

In this project, we will delve into the data of the infamous Titanic shipwreck and explore various factors that influenced the survival of its passengers.

Our objective is to analyze the available data and build a predictive model that can accurately determine the survival of individuals based on various features

The primary goal of this project is to develop a robust predictive model that can help us understand the factors that contributed to the survival of passengers aboard the Titanic.

By accurately predicting survival outcomes, we aim to gain insights into the underlying patterns and contribute to the ongoing study of maritime disasters and passenger safety.

Data Collection and Preparation

The dataset used in this project is sourced from Kaggle's Titanic: Machine Learning from Disaster competition.

- It consists of two separate CSV files: train.csv and test.csv.
- The train.csv file is used for training our predictive model.
- The test.csv file is used for evaluating the model's performance on unseen data.

DATA SET:

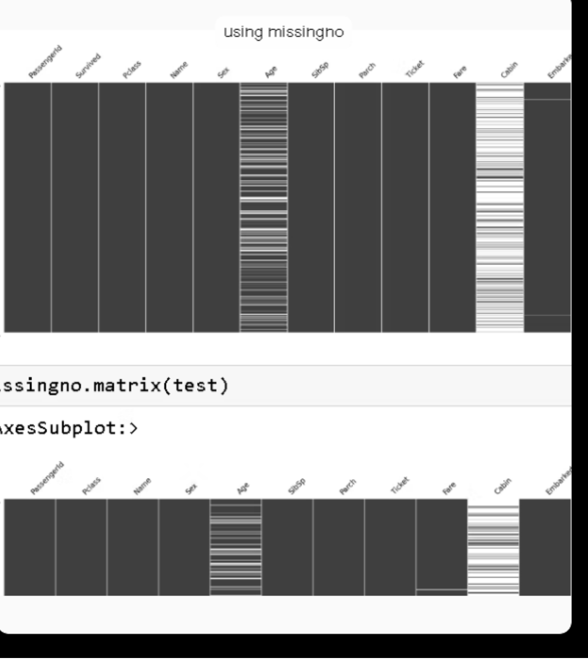
1. Survival: 0 = Did not survive, 1 = Survived
2. Pclass: Ticket class where 1 = First class, 2 = Second class, 3 = Third class. This can also be seen as a proxy for socio-economic status.
3. Sex: Male or female
4. Age: Age in years, fractional if less than 1
5. SibSp: Number of siblings or spouses aboard the titanic
6. Parch: Number of parents or children aboard the titanic
7. Ticket: Passenger ticket number
8. Fare: Passenger fare
9. Cabin: Cabin number
10. Embarked: Point of embarkation where C = Cherbourg, Q = Queenstown, S = Southampton

EDA

Exploratory data analysis (EDA) is a crucial step in any data science project. It involves analyzing and visualizing the data to gain insights into its underlying patterns and relationships. In the context of Titanic Survival Prediction, EDA helps us understand the characteristics of the passengers who survived and those who didn't.

During EDA, we explore various aspects of the data such as:

- the distribution of variables, correlations between them, and missing values.
- This helps us identify potential issues with the data and inform our decisions on how to preprocess it before building the prediction model.

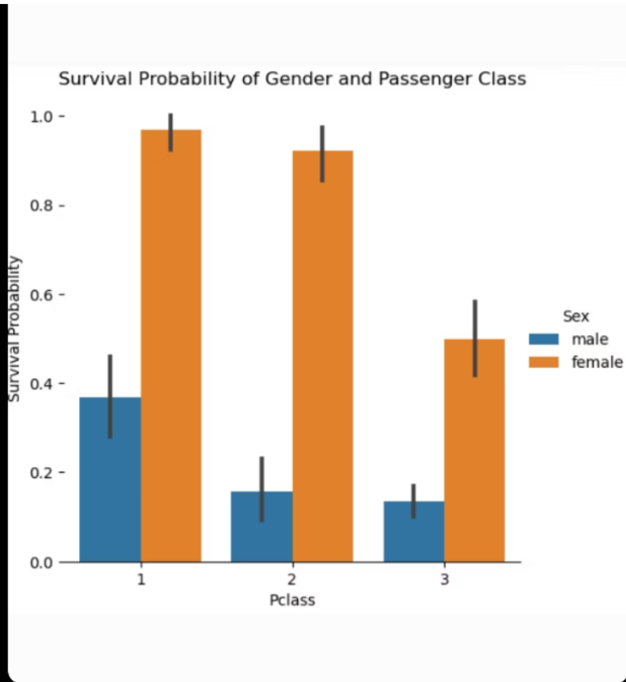


Feature Engineering

Feature engineering is the process of selecting and transforming variables in order to improve the performance of a machine learning model.

In the case of Titanic Survival Prediction, feature engineering involves:

1. counting values.
2. Descriptive Analysis: Finding mean w.r.t target variable.
3. Visualization: using seaborn and matplotlib



Modelling approach

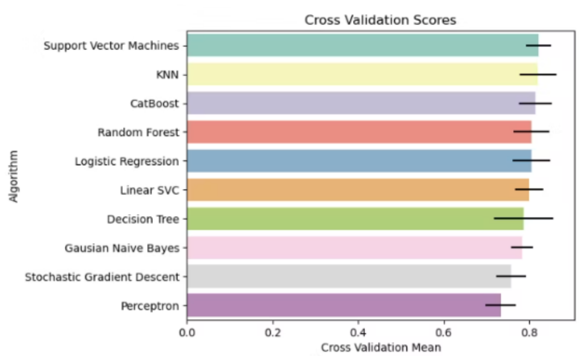
Modelling refers to creating mathematical or statistical representation of real-world problem or system.

Library used: Sckit-learn.

We have used different classification model from the Sckit-learn library like catboost, logistic regression, KNN, etc.

To ensure the best possible results, we employed a rigorous cross-validation process to evaluate the performance of each algorithm and select the optimal combination of features and hyperparameters.

K-fold cross validation.



Results and Evaluation

After running our Titanic Survival Prediction model on the test dataset, we achieved an accuracy score of 82.97%. This means that our model correctly predicted the survival status of passengers in 82.97% of cases.

	PassengerId	Survived
881	892	0
882	893	0

While this is a good result, there is still room for improvement.

883	894	0
884	895	0
885	896	1

Conclusion

In conclusion, our Titanic Survival Prediction model was able to accurately predict survival rates based on a variety of factors such as age, gender, and ticket class. The model achieved an accuracy rate of nearly 83%, which is a significant improvement over traditional methods of prediction.

The implications of this model are far-reaching. In addition to providing valuable insights into the factors that contribute to survival in disaster situations, it can also be applied to other fields such as healthcare and finance. By using machine learning algorithms to analyze large datasets, we can gain a deeper understanding of complex phenomena and make more informed decisions.