Nikhil Khanchandani

CRISP-DM Titanic Dataset Analysis Report

1. Business Understanding:The Titanic tragedy is one of the most significant maritime disasters in history, claiming the lives of over 1,500 people when it sank in 1912. This project aims to explore the factors that influenced a passenger's chances of survival by analyzing various socio-economic and personal characteristics, such as age, gender, and class. The ultimate objective of this analysis is to build a predictive model that can classify passengers as survivors or non-survivors with a target accuracy of at least 80%
2. Business Objectives: The key goal is to create a classification model that uses several metrics to evaluate its performance. These include: - Accuracy: Our aim is to reach or exceed 80% accuracy. - Precision, Recall, and F1 Score: Due to the unbalanced nature of the dataset, these metrics will ensure that the model's performance is properly measured across both classes. - ROC-AUC Score: This will help evaluate the model's ability to differentiate between survivors and non-survivors.
3. Assess Situation: This analysis uses data from the Titanic dataset available on Kaggle, which contains 891 records and features such as 'Age', 'Sex', 'Pclass', and 'Fare'. Key challenges include missing data in some fields (e.g., 'Age') and the unequal distribution between survivors and non-survivors, which could impact the model's performance. The tools utilized for this project are Python libraries such as pandas for data processing, scikit-learn for model development, and matplotlib for visualization.
4. Data Mining Goals: The main goal of this project is to predict the probability of survival for each passenger. Additionally, by examining the most important factors like 'Sex' and 'Pclass', we hope to uncover deeper insights into what drove survival outcomes, helping prioritize the most significant features for building our model.

Nikhil Khanchandani

CRISP-DM Titanic Dataset Analysis Report

5. Data Understanding: Our exploratory analysis revealed that gender was a major determinant of survival, with women having significantly higher survival rates than men. Additionally, passengers in first class had a better chance of survival compared to those in lower classes. Missing data was noted in key fields like 'Age' and 'Cabin', and we observed that critical features influencing survival were 'Sex', 'Pclass', and 'Age'.
6. Data Preparation: The dataset was cleaned and prepared for modeling by handling missing values and transforming categorical variables. For example, missing values in the 'Age' column were imputed using the median, while categorical variables like 'Sex' were encoded as binary (0 for male, 1 for female). Dummy variables were created for categorical features such as 'Embarked'. Features deemed less important, like 'Ticket' and 'Name', were removed to avoid introducing noise into the model.
7. Modeling:We employed three machine learning models to analyze the data: Logistic Regression, Decision Tree, and Random Forest. The data was split into a training set (80%) and a testing set (20%) to evaluate model performance. - Logistic Regression produced an accuracy of 81%, with decent precision and recall. - Decision Tree provided interpretable results but with less stability compared to other models. - Random Forest, as an ensemble method, delivered the best results with an accuracy of 84%, demonstrating its ability to generalize effectively to unseen data.
8. Evaluation:The Random Forest model was chosen due to its superior accuracy (84%), precision (79%), and recall (77%). This balance between precision and recall, especially in an imbalanced dataset, makes it the most suitable model for this analysis. To further improve the model, hyperparameter tuning and cross-validation could be applied.

9. Deployment: In a real-world application, this model could be deployed using web frameworks such as Flask. This would allow users to input key passenger attributes (e.g., age, sex, class) to predict their survival chances. Additionally, monitoring the model over time would help ensure it continues to make accurate predictions, with re-training being an option as more data becomes available.