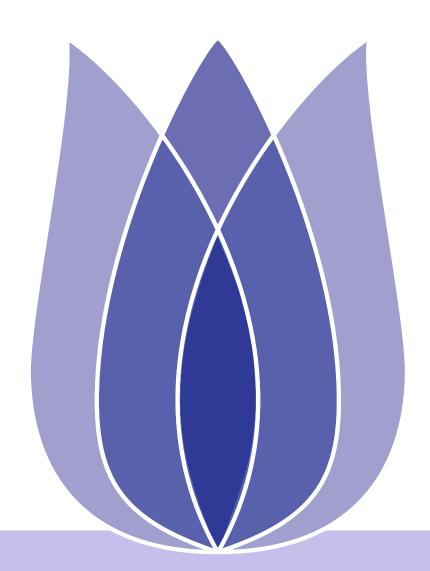
Titanic Survial Prediction

Pratikshya Parajuli

Ministry of Finance Government of Nepal

July 10, 2022





Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

Cleaning

Predictive Modeling

Conclusion

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data Cleaning

Predictive Modeling





Introduction

Overview

Exploratory Data Analysis(EDA)

Feature Engineering and Data Cleaning

Predictive Modeling

Conclusion

Introduction





Overview

Introduction

Overview

Exploratory Data Analysis(EDA)

Feature Engineering and Data Cleaning

Predictive Modeling

Conclusion

The sinking of the Titanic is one of the most infamous shipwrecks in history. This project aims to create a model that predicts which passengers survived the disaster.

- Useful features are Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare,
 Cabin, Embarked
- Target feature is Survived



Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data

Cleaning

Predictive Modeling

Conclusion

Exploratory Data Analysis(EDA)





Dataset

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data

Cleaning

Predictive Modeling

Conclusion

45	Passengerld	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	С
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

■ Age, Sex, Embarked have null values.





How many Survived?

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

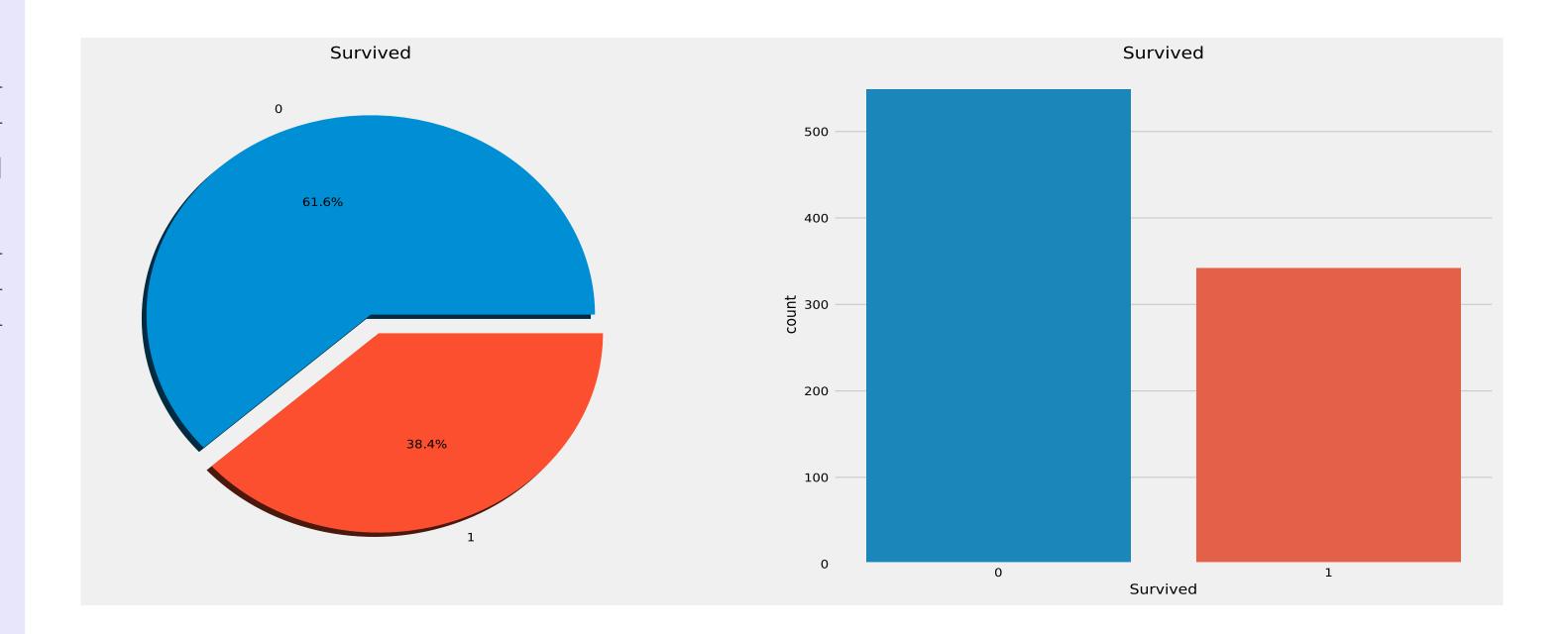
Analysis of the Features

Feature Engineering and Data

Cleaning

Predictive Modeling

Conclusion



■ We will try to check the survival rate by using the different features of the dataset. Some of the features being Sex, Port Of Embarcation, Age, etc.



Analysis of the Features

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

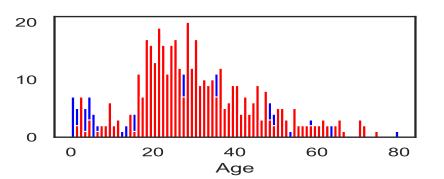
How many Survived?

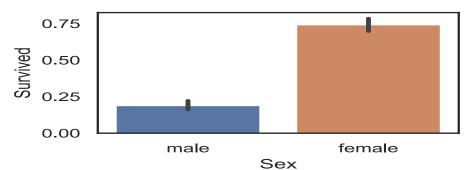
Analysis of the Features

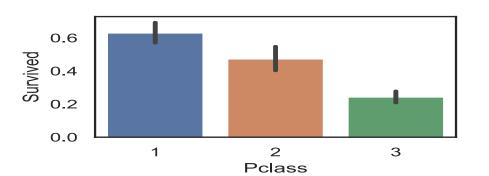
Feature Engineering and Data Cleaning

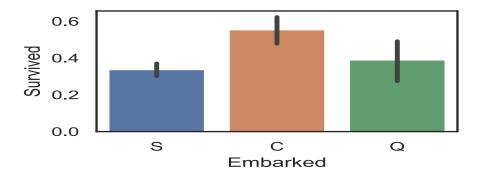
Predictive Modeling

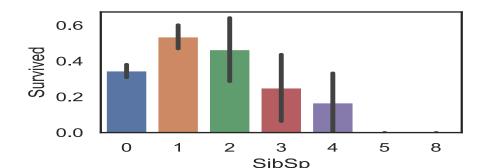
- Categorical Features in the dataset Sex, Embarked
- Ordinal Features in the dataset Pclass
- Continuous Features in the dataset Age

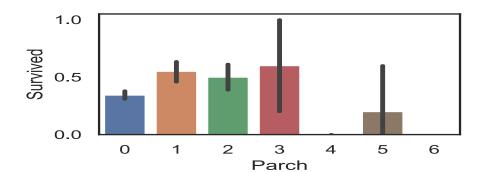


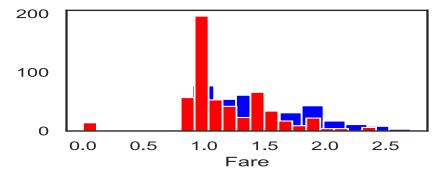














Sex - Categorical Feature

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

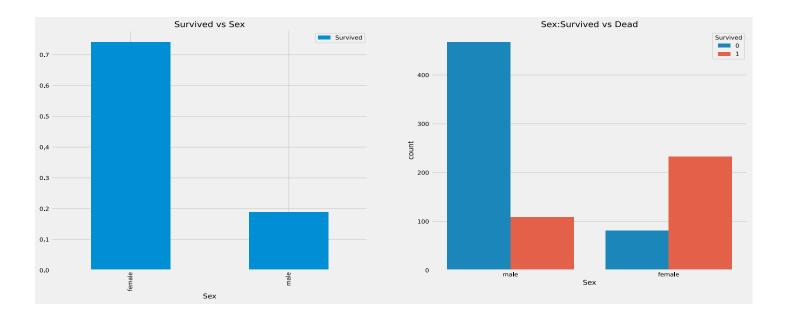
Feature Engineering and Data

Predictive Modeling

Conclusion

Table 1: Survived vs. Sex

Sex	Survived	Numbers
Female	0	81
	1	233
Male	0	468
	1	109



■ Survival rates for a women: 75 percent and men: 18-19 percent.





Pclass - Ordianal Feature

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data Cleaning

Predictive Modeling

Table 2: Numbers of Passengers by Pclass

Survive	d0	1	All
Pclass			
1	80	136	216
2	97	87	184
3	372	119	491
All	549	342	891

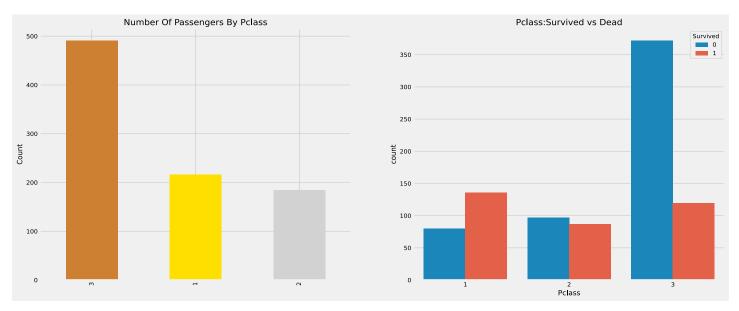


Figure 1: Pclass:Survived vs Dead





Survival rate with Sex and Pclass Together

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data Cleaning

Predictive Modeling

Table 3: Survival rate with Sex and Pclass Together

Sex	PclassSur	cv i ved	2	3	All
Female	0	3	6	72	81
	1	91	70	72	233
Male	0	77	91	300	468
	1	45	17	47	109
All		216	184	491	891

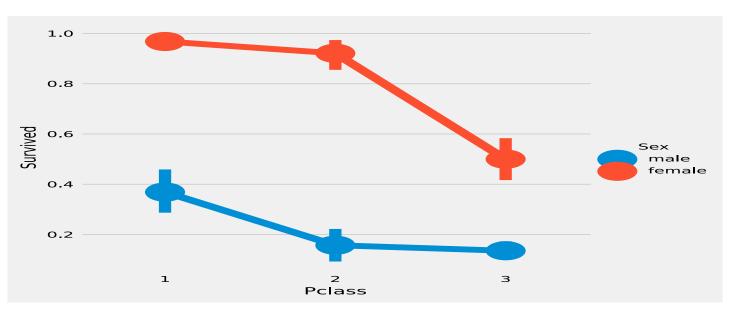


Figure 2: Survival rate with Sex and Pclass Together





Age - Continuous Feature

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data

Predictive Modeling

Conclusion

Oldest Passenger was of: 80.0 Years

Youngest Passenger was of: 0.42 Years

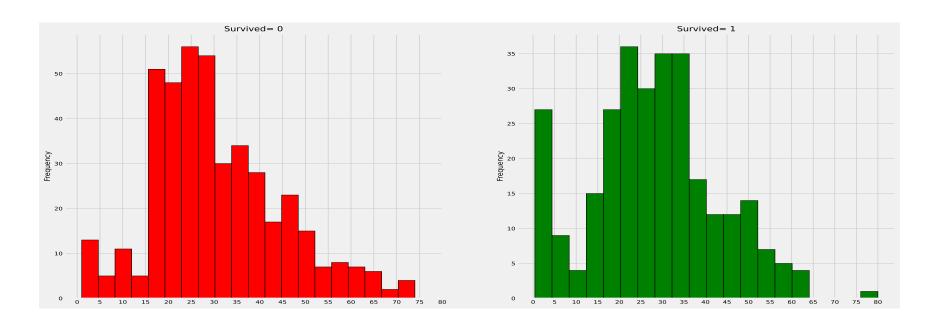


Figure 3: Survival rate with Age

Observations:

- 1)The Toddlers(age<5) were saved in large numbers(The Women and Child First Policy).
- 2) The oldest Passenger was saved (80 years).
- 3)Maximum number of deaths were in the age group of 30-40.





Embarked - Categorical Value

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data

Predictive Modeling

- 1)Maximum passenegers boarded from S. Majority of them being from Pclass3.
- 2)The Passengers from C survived.
- 3)The Embark S looks to the port from where majority of the rich people boarded. Still the chances for survival is low here.
- 4)Port Q had almost 95 percent of the passengers were from Pclass3.

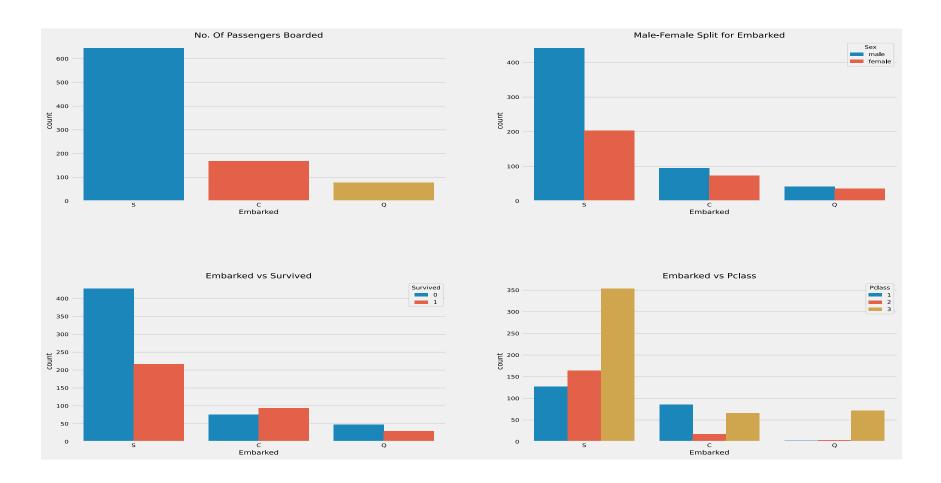


Figure 4: Survival rate with Port of Embarkation





Relation Between The Features

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data Cleaning

Predictive Modeling

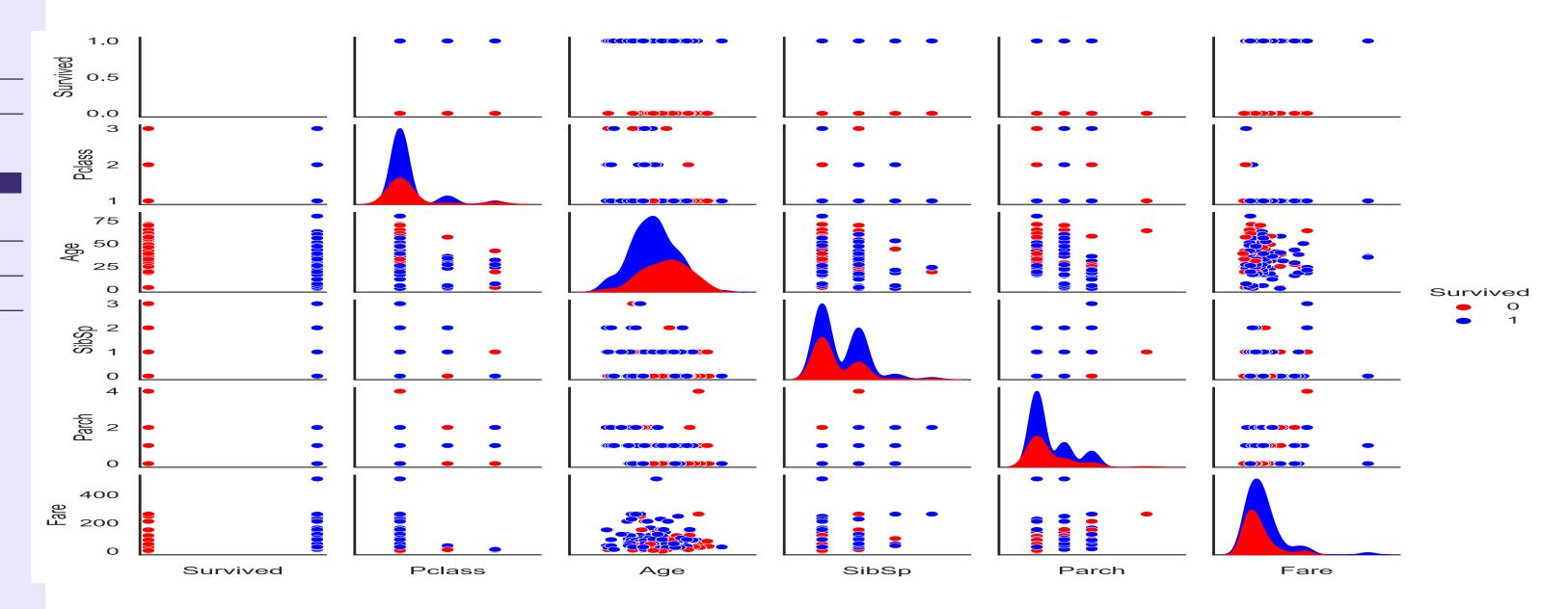


Figure 5: Relation Between The Features



Correlatoin Matrix

Overview

Introduction

Exploratory Data Analysis(EDA)

Dataset

How many Survived?

Analysis of the Features

Feature Engineering and Data

Predictive Modeling

Conclusion

The highest correlation is between SibSp and Parch i.e 0.41.



Figure 6: Interpreting the heatmap





Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data Cleaning

Removing Redundant features Correlation Matrix after Data Cleaning

Predictive Modeling

Conclusion

Feature Engineering and Data Cleaning





Converting features into suitable form for modeling

Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

Removing Redundant features Correlation Matrix after Data Cleaning

Predictive Modeling

Conclusion

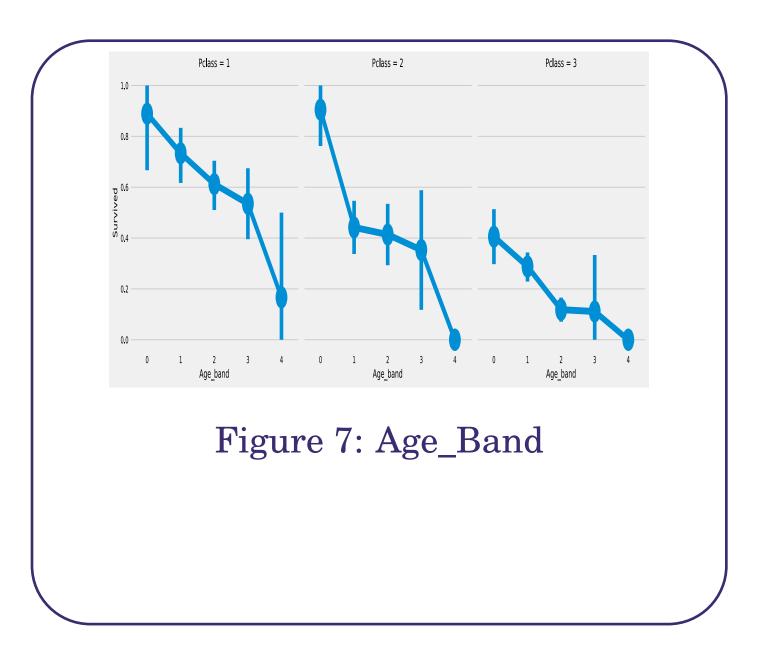
Age: Age_band

■ Family_size and Alone: Summation of Parch and SibSp

■ Fare: Fare_cat

Table 4: Age_Band

Age_band	Numbers
1	382
2	325
0	104
3	69
4	11





Removing Redundant features

Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data Cleaning

Removing Redundant features

Correlation Matrix after Data Cleaning

Predictive Modeling

- Name—> We don't need name feature as it cannot be converted into any categorical value.
- Ticket—> It is any random string that cannot be categorised.
- Fare—> We have the Fare_cat feature, so unneeded
- Cabin—> A lot of NaN values and also many passengers have multiple cabins. So this is a useless feature.
- Fare_Range-> We have the fare_cat feature.
- PassengerId—> Cannot be categorised.





Correlation Matrix after Data Cleaning

Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data
Cleaning

Cleaning

Cleaning

Removing Redundant features
Correlation Matrix after Data

Predictive Modeling

Conclusion

Positive correlation: SibSp andd Family_Size and Parch and Family_Size and Negative correlation: Alone and Family_Size

Survived	1	-0.34	0.54	-0.035	0.082	0.11	0.43	-0.11	0.017	-0.2	0.3	1.0
Pclass	-0.34	1	-0.13	0.083	0.018	0.046	-0.047	-0.31	0.066	0.14	-0.63	0.8
Sex	0.54	-0.13	1	0.11	0.25	0.12	0.63	-0.15	0.2	-0.3	0.25	0.6
SibSp	-0.035	0.083	0.11	1	0.41	-0.06	0.29	-0.26	0.89	-0.58	0.39	0.4
Parch	0.082	0.018	0.25	0.41	1	-0.079	0.31	-0.2	0.78	-0.58	0.39	
Embarked	0.11	0.046	0.12	-0.06	-0.079	1	0.12	0.024	-0.08	0.018	-0.091	0.2
Initial	0.43	-0.047	0.63	0.29	0.31	0.12	1	-0.39	0.35	-0.32	0.24	0.0
Age_band	-0.11	-0.31	-0.15	-0.26	-0.2	0.024	-0.39	1	-0.27	0.2	0.025	-0.2
Family_Size	0.017	0.066	0.2	0.89	0.78	-0.08	0.35	-0.27	1	-0.69	0.47	-0.4
Alone	-0.2	0.14	-0.3	-0.58	-0.58	0.018	-0.32	0.2	-0.69	1	-0.57	-0.4
Fare_cat	0.3	-0.63	0.25	0.39	0.39	-0.091	0.24	0.025	0.47	-0.57	1	-0.6
	Survived	Pclass	Sex	SibSp	Parch	Embarked	Initial	Age_band	Family_Size	Alone	Fare_cat	

Figure 8: Correlation Matrix after Data Cleaning





Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

Cleaning

Predictive Modeling

Prediction Accuracy

Conclusion

Predictive Modeling





Evaluation Classification Algorithms

Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

Cleaning

Predictive Modeling

Prediction Accuracy

- Logistic Regression
- Support Vector Machines (Linear and radial)
- Random Forest
- K-Nearest Neighbours
- Naive Bayes
- Decision Tree





Prediction Accuracy

Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

Predictive Modeling
Prediction Accuracy

- Split the train sample into train and test dataset
- Train Data_size : 0.7 and Test Data_size : 0.3
- Total sample size = 623; training sample size = 623, testing sample size = 268

Table 5: Accuracy Comparison of different Classifier Algorithms

	Acuracy
Radial Support Vector Machines(rbf-SVM)	0.835820895522388
Linear Support Vector Machine(linear-SVM)	0.8171641791044776
Logistic Regression	0.8134328358208955
Decision Tree	0.8059701492537313
K-Nearest Neighbours(KNN)	0.832089552238806
Gaussian Naive Bayes	0.8134328358208955
Random Forests	0.8208955223880597





Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

Cleaning

Predictive Modeling

Conclusion





Conclusion

Overview

Introduction

Exploratory Data Analysis(EDA)

Feature Engineering and Data

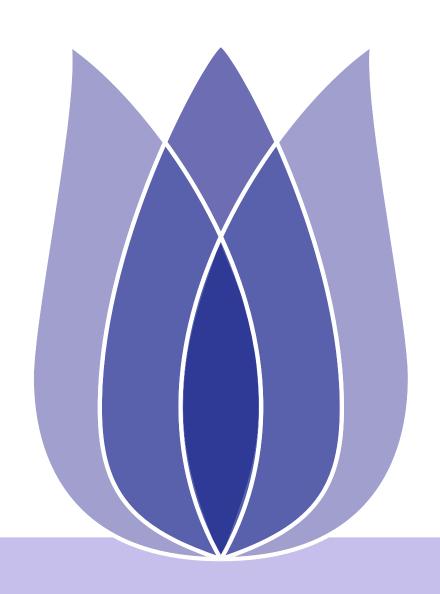
Predictive Modeling

- Basic modeling of the data
- To overcome the model variance, and get a generalized model,we can use Cross Validation
- Results can be further enhanced





Contact Information



Pratikshya Parajuli Ministry of Finance Government of Nepal



PPARAJULI@MOF.GOV.NP