

Name: Samiksha Pandey

ID: 014615237

Kaggle ID: ADS245_5237

Introduction

In this project we participated in Titanic Kaggle Competition. This project intended to create machine learning model from scratch and generate predictions for the test dataset. We followed CRISP-DM methodology in solving the problem. Next sections cover the details of each phase.

Problem Understanding

Titanic was one of the biggest disasters in our history. Few lucky people were able to survive the disaster. Problem in hand we have is to predict if the passenger survived or died based on the information available about travelers. This is a “Supervised Learning” problem. We need to develop binary classification model that predicts 0 for died and 1 for survived.

Data Understanding

Data is available as train and test dataset. Train dataset will be used to build and validate model. Test dataset is unlabeled and will be used to predict values for submission to Kaggle. While building a machine learning model it is advised to prepare training data separately from test data. Test data will not be exposed to train data to avoid any bias in the model. Train dataset has 10 columns. One target feature i.e. Survived and 9 descriptive features. Test dataset does not have target feature.

Data Exploration

First step is to explore data. Identify data quality issues and state measures to handle valid issues.

1. Data Quality Report

We start by preparing data quality report for continuous and categorical features in the dataset. In order to do so I wrote a function which gives statistical values and initial distribution of data in the columns. Age is normally distributed.

2. Data Issues

Missing Data:

Train dataset had missing values in 3 columns:

1. Age (19.8%)
2. Embarked (0.22 %)
3. Cabin (77%)

Test dataset had missing values in 2 columns:

1. Age
2. Fare

Outliers:

Age and Fare are two continuous features in both the datasets. In the box plots we can see few outliers, but all the outliers are valid data. This issue is noted and will be handled by normalization and binning techniques.

3. Data Quality Handling

For handling missing values, we have used Imputation in both train and test dataset.

Age: For imputing 177 missing ages in train dataset we used random values between mean and 1.5 standard deviation of mean. This does not change the distribution after imputation and statistical measures are also not impacted much.

Embarked: Only 2 rows had missing values in train dataset. Since this is a categorical feature, we used mode of the column 'S' and imputed missing values.

Cabin: This column had 77% missing values in train dataset. We are not removing this feature but will explore more in feature engineering. This column might be useful for preparing another feature.

Fare: Only 1 row in test dataset had missing value. This passenger was male, travelled in 3rd class, embarked from port 'S', was travelling alone, and had age 62. We used median age of passengers with similar information for imputing missing value. Also, there was no change in distribution of entire field after imputation.

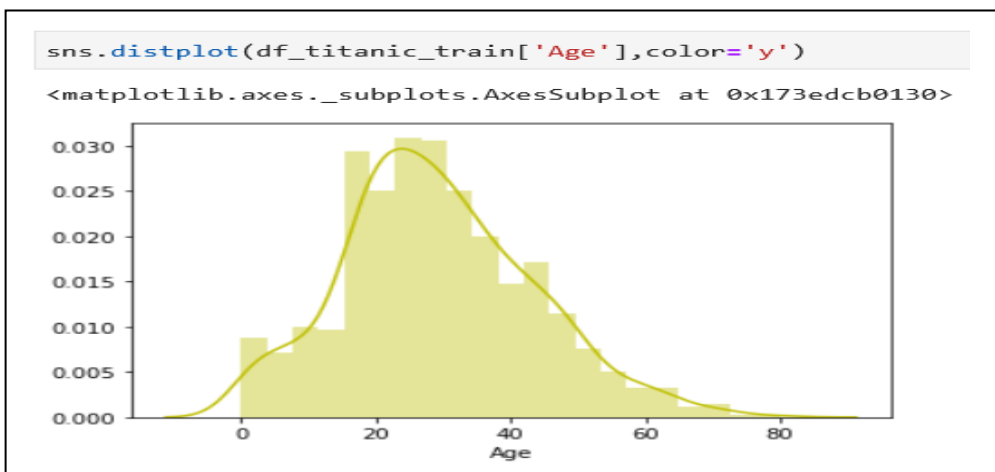
4. Data Relationships

On analysis of initial raw data, most related feature with survival is gender. Apart from that class, fare and port of embarkation have also shown some impact on chances of survival. It seemed passengers embarked from port 'S' survived most. Also, first class travelers survived most and this in turn shows passengers with 1st class had more fare, hence more fare lead to high survival chances sometimes. Age does not seem to provide any relationship with survival so far. Bot survived and died appeared to be in middle age group of 20-30 years.

All the visualizations are present in later section of this report towards end.

Data Visualization

Visualization from training dataset after handling Data Quality issues:



Rest of the visualizations are in source code part at the end of report in detail.

Data Preparation

Data is prepared and transformed for model. We start with train data and identified data transformations that will be performed. Similar data preparation is done for test dataset before making predictions using model.

Feature Engineering

Dataset has mixed features which cannot be categorized as continuous or categorical. These features are “Names, Ticket and Cabin”. These features have mixed numeric and text continuous data. Using these features directly is not feasible in the model. We explored these features to utilize in our models. Below Features are created from these columns:

1. Prefix/Prefix_Group (Using Name):
On getting closer look at this feature, we can see titles / prefixes are present in the data. We created new feature by selecting title value from Name. After selecting title, it is further grouped into Prefix_Group.
2. Deck (using Cabin):
Initially we had 77% missing values in Cabin feature. Instead of discarding it we tried to explore it. It seemed to be combination of deck and room number. Mostly, 1st class passengers have cabin numbers associated. 3rd class passengers do not have cabin. We selected first letter from cabin to prepare deck column. Missing values were replaced with ‘U’ value.
3. Travelling with/Family_Group (Using Sibsp+Prach):
Sibsp column had number of siblings or children with the passenger. Prach had number of spouses or parents with the passenger. Travelling_with column total people passenger is travelling with. Further using this new column “Family_Group”, passengers are categorized as travelling alone, in small group, in middle group or in large groups.
4. Age_group (Using Age):
Further Age is categorized into 5 age groups in this column.

Normalization

We had outliers in Fare column and Age columns. Fare column will be used in KNN model. Hence it is advised to use scaled data. Fare column is normalized using min and max normalization technique. Small function was written to achieve this.

Categorical to Numerical

Finally, all the existing and new features were converted to numerical. Sex, Embarked, Family_Group, Age_Group, Deck and Prefix_Group were converted to numerical by changing categories to values between 0-9. **Feature Selection: Final Heat map developed showed high correlation of Pclass, Sex, Fare, Embarked, Prefix_Group and Deck with survived column.**

Modeling

KNN model was developed from scratch for this process. KNN was selected as we do not have large dataset and features are converted to numerical data. KNN is easier to implement and this particular use case can be benefitted by utilizing similarity in passengers who survived to predict survival of unlabeled passenger.

Pseudo Code

1. Get Train data
2. Get Test Data or query instances
3. Require K value (nearest neighbors for prediction)
4. Iterate over test data
 - a. For each row in test data get Euclidean distance with every row in train data
 - b. Get K nearest neighbor
 - c. Make prediction using majority target level in k nearest neighbors.

Validation

To validate the model train dataset is manually split into train and test arrays. 580 rows in train array and 311 in test data. This model is validated using F1, accuracy, precision and recall for each label. Function for confusion matrix is also written. Code written is validated using different k values and k value with highest scores is selected for deployment. K=5 had highest accuracy of 85% and F1 score of 78%.

Deployment

Once our model is ready, we deploy this model for test data for making predictions. Data transformation identified for training data is applied to test data and test data is passed to KNN model with K=5.


Final predictions are saved to csv file for Kaggle submission.

Kaggle Ranking

Kaggle Score: 0.78947

Below are screenshots of Kaggle Ranking as per date **28/11/2020**.

1794	Zhenhao Wang		0.78947	10	311
1795	Mutesasra denis		0.78947	13	2h
1796	ADS245_5237		0.78947	3	17m
Your Best Entry					



ADS245_5237
MSDA Student at San Jose State University
San Jose, California, United States
Joined 9 months ago · last seen in the past day

[Home](#) [Competitions \(1\)](#) [Datasets](#) [Notebooks](#) [Discussion](#) [Organizations](#)

Competitions
Novice

Unranked

0 0 0

[Titanic: Machine...](#)
Ongoing
Top 10%

1,796th
of 18095

Datasets
Novice

Unranked

0 0 0

Notebooks
Novice

Unranked

0

Submission File:

