Titanic Survival Analysis

This report talks about predicting the survival chances of the Titanic Passengers using machine learning models with the help of Skikit-learn and TensorFlow(Keras and Estimator APIs). The code is executing using Google Colaboratory's Jupiter Notebook. The dataset used is 'titanic.csv'.

Importing essential libraries such as:

- Seaborn and Matplotlib are data visualization libraries
- Pandas is used for data manipulation and analysis
- NumPy is used for working with multi-dimensional array object
- Skikit-learn is used analyze machine learning models
- Tensorflow is used with keras and estimator APIs.

Understanding the data: After the data is loaded to Google Colab, I explored to understand the nature, structure and type of the data in the dataset.

- There are about 1304 rows with 14 columns.
- Attributes: survival, pclass, name, sex, age, sibsp, parch, ticket, fare, cabin, embarked, boat, body, home.dest
- There are variables in the form of objects that needs to be converted to float or int for analysis

Preprocess the data: There were five main steps in data pre-processing:

1. Dropping unnecessary columns namely name, ticket ID, cabin, boat, body, home.dest.
2. Converting categorical fields like sex and embarked to numerical
3. Dropping rows with 'NA' values in otherwise useful columns.
4. Reordering the columns to make easy to split and process. The independent attributes are 'pclass', 'sex', 'age', 'fare', 'sibsp', 'parch', 'embarked' and these attributes are used in prediction of the 'survival' of the passenger.
5. Split the dataset for training and testing with a ratio of 80:20

Observations made on the data

- Age: Passenger age ranges from 0-80. Most passengers belong to age range for 15-50.
- Fare: Most of the passengers have paid fare less than 100 pounds and the maximum fare amount is above 500 for first class passengers.
- Sex: Count of women passengers is less than males.
- Survival: Children, teens and women have highest chances of survival.
- Count of passengers in the third class is high but chances of survival are greater for passengers in the first class.
- Passengers with siblings, spouse, parents and children have higher chances of survival in first class and second class.
- Based on the location the passenger has boarded the ship, people from location Cherbourg have higher chances of survival.

Skikit-Learn library for predicting survival of the passenger.

Using supervised machine learning models for classification from Skikit-Learn, I am trying to predict the survival chances of a passenger. The models used are **Linear model's Logistic**

Classifier, K- Nearest Neighbors Classifier model, Gaussian Naïve Bayes Classifier model, Decision Tree Classifier model and Random Forest Classifier model. As per the results from the model, **Random Forest Classifier model has the highest training accuracy of 97.9%** , followed by **Decision Tree classifier model with 96.53%.** The results were obtained for the training dataset.

A confusion matrix is created in order to evaluate the performance of all the models. The confusion matrix summarizes the result in the matrix, indicating how many passengers were correctly and incorrectly classified by the model for the testing data. The accuracy for the testing data is calculated considering true positives, true negatives, false positive and false negative values obtained from the matrix. As per the results**, K- Nearest Neighbors Classifier** model shows a higher **testing accuracy of 82%** compared to other models.

As per the training results, Random Forest has better outcomes. Based on this model we can obtain feature importances. Feature importance indicates how valuable is each feature in the construction of the model. The more an attribute is used to make decisions in the model, the higher is the relative importance for that particular attribute. As per the results obtained, **age , fare and sex have the highest importances.**

The model was used to predict the survival chances of Jack and Rose. As per the prediction results, **Rose survives and Jack does not.**

TensorFlow for predicting survival of the passenger using the TensorFlow's high level API with keras.

The model is built by assembling layers, so that a model is now a graph of layers. The most common types of such model are **tf.keras.Sequential**. The first step is to create a fully connected network and train the model. The data is fed to the model after breaking up the entire training dataset into batches. When all the batches are fed exactly once, then it completes an epoch. By repeating this process multiple times increases the success of training the model. With multiple epoch the accuracy increases and with a certain number of epochs the accuracy remains almost the same. The results show that with epoch=50 the **accuracy is about 75% and it increases close to 80%** with epoch=70 and anything greater than 70 the accuracy remains the same.

The sequential model can be further used to predict the survival chances of Jack and Rose. As per the prediction results**, Rose survives with 96% probability**. **Jack does not survive** because the prediction **probability is 17%.**

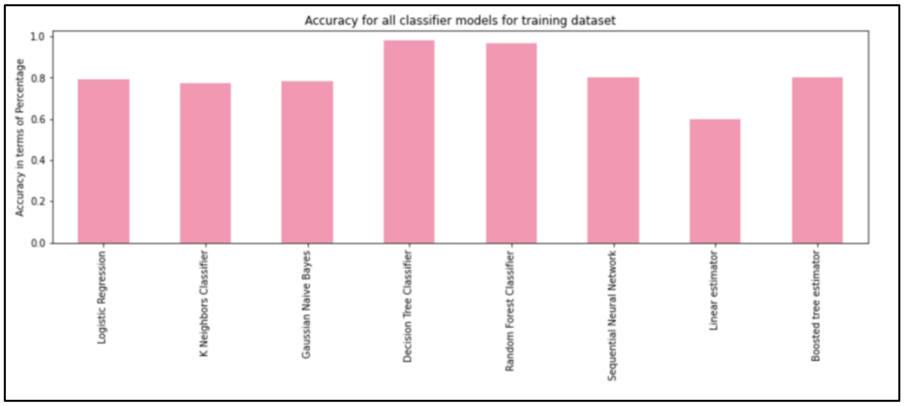
TensorFlow for predicting survival of the passenger using the TensorFlow's high level API with Estimators.

The models used here are **Linear classifier and Boosted tree.** As per the results, the accuracy of the **Boosted tree estimator is 75%** which is greater than **linear estimator's accuracy of 69%** . Based on individual performance, the linear estimator performs exactly of what was expected as a baseline accuracy, whereas boosted tree exceeds the baseline accuracy and performs significantly better.

Feature Importances for Boosted Tree Classifier: As per the training and testing accuracy results, boosted tree has better outcomes. The results show that **sex, pclass and age have the highest importances.** These attributes determine the survival chances of the passenger.

Unlike Skikit-Learn, TensorFlow provides results in the form of probabilities which can be seen as the percentage of survival for the passenger. Since boosted trees have better outcomes compared to linear estimator, the model was used to predict the outcomes for testing data. As per the analysis, there are greater number of people who did not survive and there are significantly higher number of people on the lower probability scale who have very less chance of survival. Between 80% to 97% there are very few passengers.

Conclusion:



- Of the 8 models that I used for analysis **Decision tress and Random Forest Classifier have the highest accuracy.** Logistic regression, K nearest Neighbors , Gaussian Naïve Bayes, boosted tress estimator and Sequential neural network model have almost the same accuracy and the lowest value for accuracy is for a Linear estimator model.
- The **most important attributes** which influenced the accuracy the most are **age, fare, pclass and sex.**