# Abstract

Maritime shipping has now become an important part of worldwide trade as a result of globalization. As a result, the reliability of cargo vessel shipping company has become an important element in the shipping business, as inconsistencies may cause additional delays farther down the supply chain, raising the entire cost of shipping goods or may destroy the shipping goods. In this project, we are using different built-in machine learning models for the prediction of best suitable service provider based on the different features of Cargo Shipment. Here, we also proposed a new model to more accurately predict the suitable service provider.

# Dataset Description

## Introduction:

The dataset of shipping cargos with shipment details was downloaded from Kaggle. The dataset contained the several features related to the cargos with the name of the service provider company. The downloaded dataset was used in this project for the prediction of best suitable service provider company on the basis of cargo features.

The original Dataset contains the following number of rows and Columns:

|  |  |  |
| --- | --- | --- |
| **Voyage Dataset** | **Number of Rows in Dataset:** | 21589 |
| **Number of Variables in Dataset:** | 15 |

# Dataset – Summary of Attributes

The dataset was saved in excel file and the variable detail of the dataset is available in the following table. The table described the variable name, description of variable, and type of variable.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variable Name | Description | Type |
| 1 | Vessel Name | Name of the Vessel | Categorical |
| 2 | Vessel Type | Type of the Vessel (On the basis of carrying goods) | Categorical |
| 3 | Vessel Capacity | Maximum Size or capacity of vessel of carrying goods | Numeric |
| 4 | Load Date | The date on which vessel was loaded with goods | Date |
| 5 | Category of Trade | Category of goods that was trade. | Categorical |
| 6 | Cargo Description | Brief description about the cargo. | Text |
| 7 | Volume/ Amount | Size or volume of the goods in the cargo. | Numeric |
| 8 | Volume Type | Unit of measuring the volume of the goods in Cargo | Categorical |
| 9 | Load Port | Name of the port from the cargo was loaded. | Categorical |
| 10 | Discharge Port | Destination Port of cargo | Categorical |
| 11 | Dangerous Goods | Flag for indicating that the cargo is containing the Dangerous goods or not? | Boolean |
| 12 | Organization | Name of the Organization that provide or trade the cargo from load port to destination port. | Categorical |
| 13 | Application Number | Identity number provided by organization | Numeric |
| 14 | License Number | License number of the organization. | Numeric |
| 15 | Voyage Number | Identity Number of the Cargo/Vessel | Numeric |

# Data Cleaning

## Remove Identical Features

As the dataset have the total 15 features, but there are several features that are based on identity and does not correlate with the other attributes. In other words, these attributes are unique identity n umber and did not affect the any other features. So, all the features that have static effect were removed from the dataset. The status of all the features after this procedure is presented in following table.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Status | Variable Name | Status |
| Vessel Name | Removed | Load Port | Keep |
| Vessel Type | Removed | Discharge Port | Keep |
| Vessel Capacity | Removed | Dangerous Goods | Keep |
| Load Date | Removed | Organization | Keep |
| Category of Trade | Keep | Application Number | Removed |
| Cargo Description | Removed | License Number | Removed |
| Volume/ Amount | Keep | Voyage Number | Removed |
| Volume Type | Keep |  |  |

## Attribute data types and Statistics

The Voyage report dataset has the total 15 attributes. The datatype of each attribute with the elementary statistics is following:

|  |  |  |  |
| --- | --- | --- | --- |
| Voyage Dataset | | | |
|  | **Variable Name** | **Type** | Statistics |
| 1 | Vessel Name | Categorical |  |
| 2 | Vessel Type | Categorical |  |
| 3 | Vessel Capacity | Numeric |  |
| 4 | Load Date | Date |  |
| 5 | Category of Trade | Categorical |  |
| 6 | Cargo Description | Text |  |
| 7 | Volume/ Amount | Numeric |  |
| 8 | Volume Type | Categorical |  |
| 9 | Load Port | Categorical |  |
| 10 | Discharge Port | Categorical |  |
| 11 | Dangerous Goods | Boolean |  |
| 12 | Organization | Categorical |  |
| 13 | Application Number | Numeric |  |
| 14 | License Number | Numeric |  |
| 15 | Voyage Number | Numeric |  |

## Target Variable Description

After the cleaning of the dataset, the “Organization” feature was set as target variable. The target variable contains the different number of samples for 81 unique organizations that provide cargo services. The target variable was used for the prediction of best suitable service provider based on the other attributes in dataset. The comprehensive information of Selected features and target variable is available in below table.

|  |  |
| --- | --- |
| Features | Category of Trade, Volume Type, Load Port, Discharge Port, Dangerous Goods |
| Target | Organization |

As the target variable in Voyage dataset is in Categorical form that require the label encoding. Label encoding is a technique of converting string categorical value into numerical value.

## Identification of missing values and remove outliers

Missing value in the tabular data usually contain the NAN, undefined and missing values. All the features were pass to the filter to check the NAN values and fill it with mean value of that feature. But the dataset was well structured and clean enough that the filter did not find any missing value.

# Quality of Data

The final set of data contains 6 columns for Voyage dataset. The total numbers of observations (rows) remaining in dataset is 21589.

Columns have been checked for the existence of extreme values (outliers) and actions have been taken  
to remove these observations out of the dataset. This decision has not been taken lightly, while outliers  
may indicate something scientifically interesting. In this scenario, it is believed the presence of outliers  
may hinder the performance of the classifiers.

Furthermore, columns have also been checked for missing values. In particular the removal of missing  
values started by taking into consideration priority columns such as the target variable column  
precipitation type. Since the absence of a value within that column will render this column as impractical  
a decision has been taken to remove these observations out of the dataset.  
Additionally, columns have been checked for content validity by removing observations where incorrect  
data has been identified. Repeated: not repeated. The above section describes the detail of preprocessing and this paragraph indicate the features of dataset to tell the quality of dataset.

# Train Test Split of Dataset

Furthermore, the original dataset was split into training, and testing set with the ratio of 70%, 10% and 20% respectively. For the division of the dataset, the train test split function of python library (scikit-learn) was used that shuffle the dataset and randomly split the samples in each set. This approach was used to split the dataset with balance samples of each class. The training and validation set was usually used during the training of the model while the testing set was used for the testing of the models. The total number of variables and samples in training, testing and validation set is given below:

|  |  |  |
| --- | --- | --- |
|  | **No. of Samples** | **No of Columns** |
| **Training Set** | 21271 | 6 |
| **Testing Set** | 4318 | 6 |

# Machine Learning Model Implementation

After the cleaning and preprocessing of the dataset, identify the machine learning problem for target variable. As the target variable in Voyage datasets have the Categorical value that indicate that the prediction of target variable made possible by classification models. Hence, different machine learning models for target variable prediction were chosen.

The list of all chosen classification models is below:

* Random Forest
* XG Boost
* SVM
* KNN
* Proposed Model

After the training of the models, all models were tested on test set of the dataset to evaluate the performance of trained models. Below are some evaluation measures that were used to evaluate the performance of the model.



**Accuracy -** Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

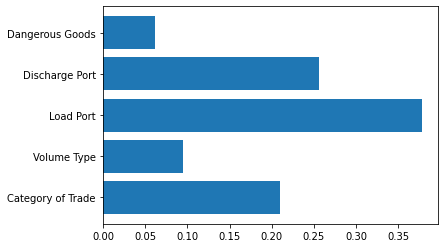
**Precision -** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

**Recall (Sensitivity) -** Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

**F1 score -** F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.

## Random Forest

The random-forest is the ensemble-learning technique for the classification, regression, and different tasks which works by building a large number of decisions-trees during the learning. For the classifications tasks, randomized forest's result is class chosen by majority of the trees. In this project, we used the Random Forest for the prediction of best service provider. The training set was used to train the model during training process. Random Forest model was trained with the default parameters or by tunning the hyper parameters. In the hyper parameter tunning, n-estimator, max-depth and random-state were tunned with different values while rest of the parameters were used with default values. The importance of features for the Prediction of service provider using Random Forest model is showed below:



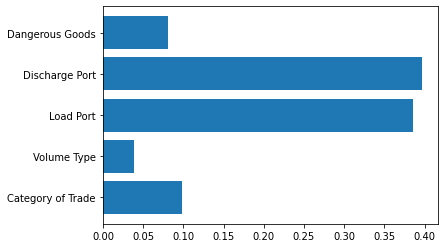
The total five input features of 17271 samples in training set were passed for the training of the model while 17271 labels of training samples were also passed as output variable. Random Forest model firstly train with the default parameters and showed the 0.7923 accuracy score. The accuracy score was calculated on the test set with the total 4318 samples.

Later, random was trained with the different values of n-estimator, max-depth and random-state and it showed the best .7925% accuracy on test set with estimators:50, max-depth:11 and random-state:0. The description of each tuned hyper parameter is below.

* n-estimators: The number of trees in the forest.
* max-depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min-samples-split samples.
* Random-state: Controls both the randomness of the bootstrapping of the samples used when building trees.

## XGBoost

XGBoost is the machine-learning method that can be employed for a variety of applications, including the regression and classifications. The XGBoost tree model is constructed in the same phase-wise manner as some other boosting approaches, but it differs in that it allows optimization of any computable error function. In this project, we used the XGBoost for the prediction of best service provider. The training set was used to train the model during training process. Gradient boosting model was trained with the default parameters or by tunning the hyper parameters. In the hyper parameter tunning, n-estimator, learning-rate and max-depth were tune with different values while rest of the parameters were used with default values. The importance of features for the Prediction of service provider using Random Forest model is showed below:



The total five input features of 17271 samples in training set were passed for the training of the model while 17271 labels of training samples were also passed as output variable. Gradient boosting model firstly train with the default parameters and showed the 0.7825 accuracy score. The accuracy score was calculated on the test set with the total 4318 samples.

Later, random was trained with the different values of n-estimator and it showed the best 0.7948 accuracy on test set with 10 estimators, maximum-depth:5 and 0.2 learning rate. The description of each tuned hyper parameter is below.

* n-estimators: Number of boosting rounds.
* max-depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min-samples-split samples.
* Learning-rate: Boosting learning rate

## SVM

Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. In this project, we used the SVM model for the prediction of best service providers. The data was split into training and testing set with 17271 and 4378 samples respectively. The training set was used for the training of the model while the test set was used for the testing of SVM model. The SVM model was firstly train with the default parameters and RBF kernel on training data. After the complete training of the model, the SVM model showed the 0.6721% accuracy on the 4318 samples of test set.

Later, the hyper parameter tunning technique was applied on the SVM model. In hyper parameter tunning, the different values of kernel, gamma and C were used to trained the model. The SVM model showed the highest accuracy (0.7186%) with RBF Kernel, gamma:1, and C:0.5 value. The description of each tuned hyper parameter is below.

* Kernel: Specifies the kernel type to be used in the algorithm.
* gamma: Kernel coefficient for ‘RBF’, ‘poly’ and ‘sigmoid’.
* C: Regularization parameter. The strength of the regularization is inversely proportional to C.

## KNN

This algorithm is used to solve the classification model problems. K-nearest neighbor or KNN algorithm basically creates an imaginary boundary to classify the data. When new data points come in, the algorithm will try to predict that to the nearest of the boundary line. In this project, we used the KNN model for the prediction of best service providers. The data was split into training and testing set with 17271 and 4378 samples respectively. The training set was used for the training of the model while the test set was used to validate the KNN model. The KNN model was firstly train with the default parameters on training data. After the complete training of the model, the KNN model showed the 0.7564% accuracy on the 4318 samples of test set.

Later, the hyper parameter tunning technique was applied on the KNN model. In hyper parameter tunning, the different values of n-neighbors, P and leaf-size were used to trained the model. The KNN model showed the highest accuracy (0.7754%) with n-neighbors:6, P:1 and leaf-size:3. The description of each tuned hyper parameter is below.

* N- neighbors: Number of neighbors to use by default for k-neighbors queries.
* leaf-size: Leaf size passed to Ball Tree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree.
* P: Power parameter for the Minkowski metric

## Customize Model

Lastly, we proposed our own model for the prediction of best service provider. Our proposed customized model is a combination of three built-in machine learning models following by the voting classifier. The three built-in machine learning models were including the Random Forest, XGBoost and KNN. The models were developed by using the Sklearn library of python. Firstly, Random Forest, XGBoost and KNN model was import and initialize to predict the best shipping service provider. The hyper parameter was tune to the same values as used for the individual models in above section. The prediction result of all the models were passed to the voting classifier. The purpose of voting classifier is to take decision on majority basis. In our case, voting classifier receive the prediction result of Random Forest, XGBoost, and KNN. Voting Classifier predict the shipping service provider base on the majority vote by Random Forest, XGBoost and KNN.

In the Sklearn library, all the machine learning models were initialized with tunned parameter values. Then the voting classifier was also initialized and the machine learning models were passed to this classifier as parameter values. The equal weight was used for the output of each model. The data was split into training and testing set with 17271 and 4378 samples respectively. The training set was used for the training of the voting classifier while the test set was used to validate the model. The voting model was train with the default parameters and equal weights on training data. After the complete training of the model, the voting classifier showed the 0.7962% accuracy on the 4318 samples of test set.

# Testing

All the models were test against the testing set of the dataset. The testing set was passes to all the trained model for making predictions. The testing set was used to calculate the score of each evaluation measure for every model by using the evaluation equations (describe above).

**Testing Set –** Testing set was based on the total six attributes like the training set. The Organization variable in testing set was set as target and rest of the attributes were selected for the testing of trained model. The testing set have the five input features and one target label of 4318 number of samples. The details of testing set are given in below table.

|  |  |  |
| --- | --- | --- |
| **Voyage Dataset** | Number of Samples in testing set | 4318 |
| Total attributes in testing set | 6 |
| Target Variable in testing set | Organization |

**Model Prediction -** Next the testing set was used for the prediction of service provider by each trained model. All the evaluation measures including the accuracy, precision, recall, and F1 score were calculated for testing sets. All the evaluation measures were calculated by using the built functions in Sklearn library of python.

**Result Analysis –** The selected evaluation measures were calculated on testing set to judge the best trained model for the prediction of best service provider. Among the built-in machine learning model, RF model perform well in term of accuracy as it showed the highest accuracy. But the proposed model crossover all the model and outer perform by all other trained models. The customized model showed the 0.7960% accuracy for prediction of best service provider. The score of evaluation measures of all model is also shown in below table. The comparative overview of all models’ accuracies is also presented in below figure.

Table 1: Evaluation Measures report of testing set with default parameters models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| Random Forest | 0.7923 | 0.7969 | 0.7923 | 0.7665 |
| XG Boost | 0.7825 | 0.7832 | 0.7825 | 0.7527 |
| SVM | 0.6721 | 0.6194 | 0.6721 | 0.6237 |
| KNN | 0.7564 | 0.7699 | 0.7564 | 0.7422 |
| customize | 0.7948 | 0.7804 | 0.7948 | 0.7650 |

Table 2: Evaluation Measures report of testing set with tunned parameters models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 |
| Random Forest | 0.7774 | 0.7594 | 0.7774 | 0.7433 |
| XG Boost | 0.7730 | 0.7700 | 0.7730 | 0.7447 |
| SVM | 0.7186 | 0.6930 | 0.7186 | 0.6819 |
| KNN | 0.7754 | 0.7927 | 0.7754 | 0.7523 |
| customize | 0.7962 | 0.7827 | 0.7962 | 0.7673 |

# Predict Best 4 Service Provider

Lastly, we predict the best four shipping service provider by using our customized model. The sample was collected by getting the input from user. Convert the sample feature values into feature vector and scale the feature vector to normalized it. After the preparing of user collected feature vector, the feature vector was pass to our customized model. The probability of each class for feature vector was predicted by our customized model. Customized model predicts the probability of happing the target label for each class. The probabilities of all classes were stored and then sorted in list. Furthermore, top 4 highest probabilities from the sorted list were retrieved as the label of best shipping service providers. Lastly the numeric labels of best shipping service providers were converted into string label by using label encoded dictionary.