RIDE FARE CLASSIFICATION

CS5621 - Machine Learning

MSc in Computer Science – University of Moratuwa

209381P – T.P.N. Silva

Kaggle User Name : priyanjalasilva

Kaggle Display Name : Priyanjala Silva

Best Result Score : 0.97665

Public Leaderboard Rank : 40

Git Repository : https://github.com/PRISLK/ML_Classification.git

1 - Provide a brief introduction to your solution.

In preprocessing, data rows with missing values were replaced with the average.

Decision Tree(ID3), SVC and Random Forest Classifier were used as the main classifiers to classify the preprocessed data. Following classes of the open source library sklearn were used to implement each of the classifier respectively.

- sklearn.tree.DecisionTreeClassifier
- sklearn.svm.SVC
- sklearn.ensemble.RandomForestClassifier

2 - Comment on feature engineering techniques you used.

- Imputation: Appearing of missing values in a data set can lead to errors. Therefore during preprocessing of data these missing values should be handled. Rather than dropping the rows with missing data, they are replaced with the average. This helps to preserve the data size.
- Handling Outliers: The machine learning models Decision Tree (ID3), SVC and Random Forest are not sensitive to outliers. Because of that reason there was no need to handle them.

3 - What classification techniques were tried?

Following classification techniques were used in the implementation.

- Decision Tree (ID3): It is a supervised learning classification technique. In this implementation ID3 algorithm is used. The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking.
- SVC: Support Vector Classification is another implementation of Support Vector Machines which is a supervised machine learning technique that can be used for classification.
- Random Forest: It is a supervised learning algorithm that uses ensemble technique for creating decision trees and trained with bagging method.

4 - What generalization techniques were used to avoid overfitting?

Overfitting occurs when the model performs well on training data but generalizes poorly to unseen data. Some of the generalization techniques used to avoid overfitting are as follows.

- Cross Validation: Splitting the dataset into k groups (k-fold cross-validation). Let one of the groups to be the testing set and the others as the training set, and repeat this process until each individual group has been used as the testing set (e.g., k repeats). Cross-validation allows all data to be eventually used for training but is also more computationally expensive than hold-out.

- Feature Selection: When having only a limited amount of training samples, each with a large number of features, should only select the most important features for training so that the model doesn't need to learn for so many features and eventually overfit. In this implementation the features such as trip_id, pickup_time, drop_time were removed.

5 - Which sampling techniques were used?

Sampling is a technique that allows to derive information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual. This is done to draw conclusions about populations from samples, and it enables us to determine a population's characteristics by directly observing only a portion (or sample) of the population.

In probability sampling, every element of the population has an equal chance of being selected. Probability sampling gives us the best chance to create a sample that is truly representative of the population.

Basic sampling technique used here is the Simple Random Sampling Technique. In Random Sampling every individual is chosen entirely by chance and each member of the population has an equal chance of being selected.

6 - Comment on any ensemble techniques used to improve results.

Bagging is a method of ensemble that combines the results of multiple models (for instance, all decision trees) to get a generalized result.

Bootstrapping is a bagging technique in which subsets of observations created from the original dataset with replacement. The size of the subsets is the same as the size of the original set. Bagging (or Bootstrap Aggregating) technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set.

Random forest is an algorithm that uses bootstrapping (Bagging algorithm). In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e. a bootstrap sample) from the training set. In addition, instead of using all the features, a random subset of features is selected, further randomizing the tree. As a result, the bias of the forest increases slightly, but due to the averaging of less correlated trees, its variance decreases, resulting in an overall better model.

Some important Parameters of Random forest used in implementation of the model are given below.

- min_samples_split: Used to define the minimum number of samples required in a leaf node before a split is attempted. If the number of samples is less than the required number, the node is not split.
- min_samples_leaf: This defines the minimum number of samples required to be at a leaf node. Smaller leaf size makes the model more prone to capturing noise in train data.
- max_depth: Random forest has multiple decision trees. This parameter defines the maximum depth of the trees.

7 - Any noteworthy observations (or conclusion drawn) from the project.

It was noted that highest public score of 0.97665 (private score : 0.97502) was obtained for the classifier Random Forest and the parameters passed had following values.

- min_samples_split=5
- min_samples_leaf=1
- n_estimators = 80
- bootstrap=True

Moreover in the dataset used Columns.trip_id, Columns.pickup_time, Columns.drop_time had been eliminated as they had no direct influence in the improvement of the classification.