# ITMD -525

# Final Project Report

## Group 12: Classification of Products


| First Name | Last Name | CWID |
|---|---|---|
| Sarin | Vellore Ravishanker | A20349372 |
| Sharan | Prakash Babu | A20356488 |

# Table of Contents

# 1. Introduction

This project studies classification methods and try to find the best model for the product classification. Machine learning models deployed in this paper include decision trees, gradient boosting model, etc. We will use percentage split as the evaluation method for prediction accuracy in order to compare between models. We also build ensemble models from the best performing classifiers to help train the model in a better performance.

## 1.1 Problem Statement

Given a dataset concerning ecommerce products with 93 features for more than 200,000 products, this project is aimed to build a predictive model that is able to distinguish products between 9 main product categories. A few selected classification learning models will be trained by the dataset that includes each product's corresponding category. In order to compare among applied models, we will evaluate the accuracy and then the model with comparatively higher accuracy will be selected.

## 1.2 Tools Used

**WEKA** - It is a widely used toolkit for machine learning and data mining originally developed at the University of Waikato in New Zealand. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for regression, classification, clustering, association rules, visualization, and data pre-processing.
**R-Programming -** R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

# 2. Dataset

- Data Source: [https://www.kaggle.com/c/otto-group-product-classification-challenge/data](https://www.kaggle.com/c/otto-group-product-classification-challenge/data)
- Number of Records: 61,878

dataset.csv

- Sample Data:

| Feature | Type | Description |
|---|---|---|
| id | numeric | anonymous id unique to a product |
| feat_1, feat_2, ...,feat_93 | numeric | various features of a product |
| target | nominal | the class of a product |

**Table 1: Description of features for products classification**

- Each instance represents single product. 93 obfuscated features are provided for the datasets, which represent counts of different events.
- The randomly selected products are to be classified into nine categories. Each target category corresponds to one of the most important product categories (like fashion, electronics, etc.).

## 2.1 Data Preprocessing/Transformation

The dataset required no major preprocessing since it had no missing values and all 93 attributes were required to classify the labels.

However, the label column was renamed to **"Category"** and all the 9 classes under the Category label where replaced with meaningful name, namely Electronics, Books,Fashion,Beauty and Personal Care, Fine Art, Gift Cards, Grocery, Household and Pet Supplies.

## 2.2 Data Exploration

- The features in the dataset are completely obfuscated; the meaning behind the 93 features and what the 9 Categories are remains unknown. We only know the features are integer counts and contain many zeros.

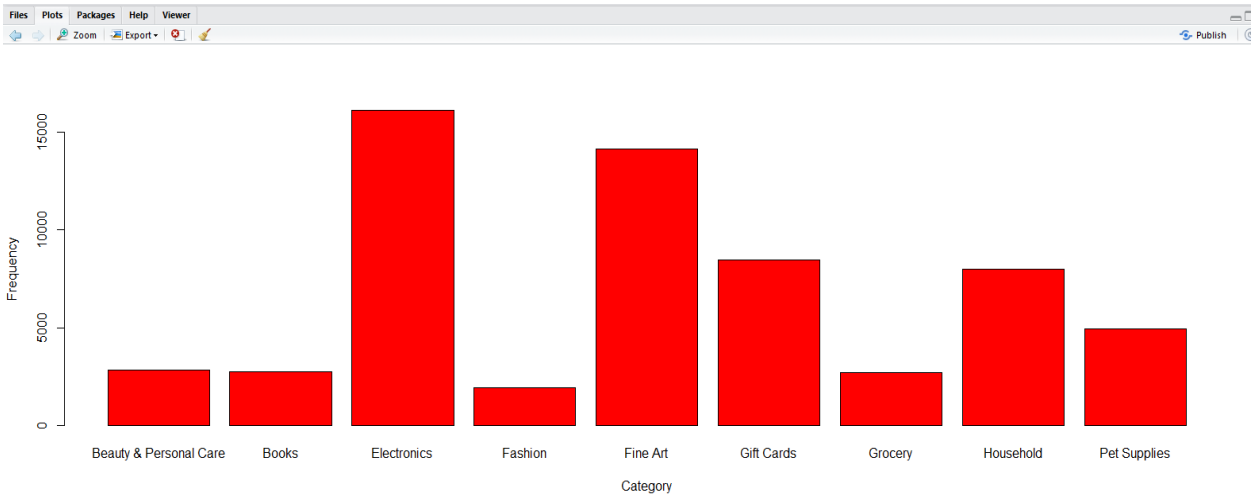


**Figure 1 – Distribution of 9 product categories**

- On sub setting the data set to 15000 records, we can see the distribution of the 9 classes in 2D space using t-SNE technique. On analyzing the data points, it can be said that the data set has outliers and not uniformly scattered.

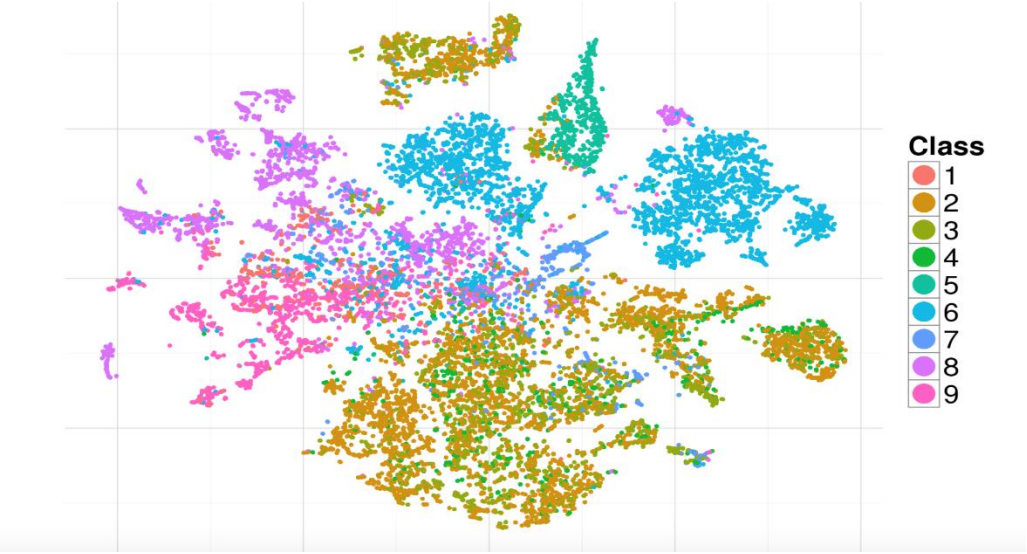


**Figure 2 – 2D Embedding of Products Data**

# 3. Approach

- Cleaning and Pre-Processing of Data
- Classified the labels using different classification algorithms
- Predicted important features from the dataset
- Compared accuracy of the classifiers in Weka and R
- Built an Ensemble model to improve classification accuracy

# 4. KDD

## 4.1. Data Processing

- The dataset required no major preprocessing since it had no missing values and all 93 attributes were required to classify the labels.
- However, the label column was renamed to **"Category"** and all the 9 classes under the Category label where replaced with meaningful name, namely Electronics, Books,Fashion,Beauty and Personal Care, Fine Art, Gift Cards, Grocery, Household and Pet Supplies.

## 4.2. Data Mining Processes

The Classification Models used for classification are:

1) K-Nearest Neighbor Classifier

2) Random Forest Algorithm
3) ADA Boost Algorithm
4) XG Boost Algorithm
5) Ensemble Model

1) **K-Nearest Neighbor Classifier**
   - K-Nearest neighbor (kNN) is a classification strategy that is an example of a "lazy learner." Unlike all the other classification algorithms "lazy learners" do not require building a model with a training set before actual use. A lazy learner like k-nearest neighbors uses the training set directly to classify an input when an input is given.
   - We run 3 iterations for different K-values (**k=9,k=15,k=25**) which yields the best accuracy on both Weka and R.

**2) Random Forest Algorithm**
   - Random Forest is one of the most widely used machine learning algorithm for classification. It can also be used for regression model (i.e. continuous target variable) but it mainly performs well on classification model (i.e. categorical target variable).
   - One of the important parameters in this algorithm is **"Number of Trees" (ntree).**
   - We run 3 iterations by tuning the value of number of trees to obtain the best accuracy on both Weka and R.

**3) Adaptive(ADA) Boost Algorithm**
   - The term 'Boosting' refers to a family of algorithms which converts weak learner to strong learners.
   - With ADA Boost, different types of learning algorithms can be used in conjunction to improve the model performance.
   - The Parameter used for tuning was "**mfinal**" in R which indicates the number of iterations for which boosting is run or number of trees used.
   - an integer, the number of iterations for which boosting is run or the number of trees to

**4) Extreme Gradient(XG) Boosting Algorithm**
   - XGBoost is an implementation of the gradient boosting algorithm, this model is often described as a black box: it works well but it is not trivial to understand how. Widely used for supervised learning problems, the training data (with multiple features) xi to predict a target variable yi. These models are made of hundreds of decision trees.
   - XGBoost is known for its fast speed and accurate predictive power, it also comes with various functions to help us understand the model
   - The Parameter used for tuning was "**nrounds**" in R which indicates the number of iterations for which boosting is run or number of trees used.

5) **Ensemble Model**

- As part of the classification of the products an Ensemble model is designed to increase the classification accuracy. The model taken in probabilities of classifying into different labels from the following models:
  - XGBoost
  - ADABoost
  - Random Forest
- Each algorithm's probabilities are given weight and based on the new probability values the labels are assigned. As part of our model the weights of different models that increased the overall classification are shown in the below diagram.

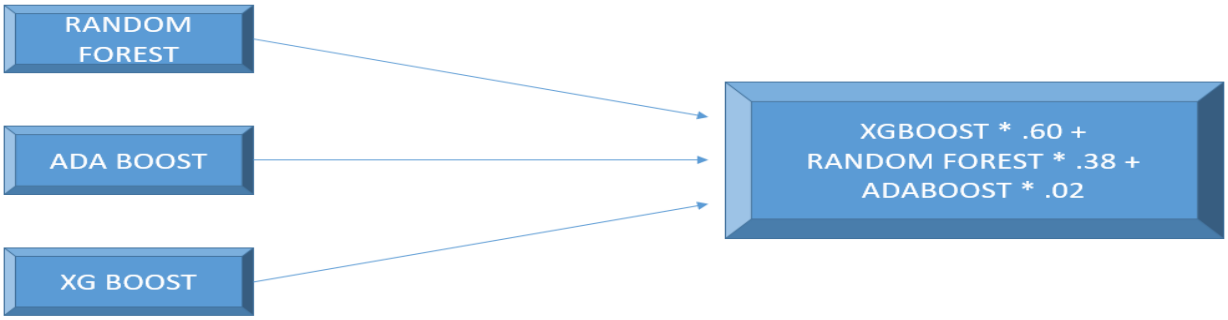


**Figure 3 - Ensemble Model**

# 5. Evaluations and Results

## 5.1. Evaluation Method

- The Evaluation method used for the classification process is percentage split of the dataset, where 60% (37,127 records) was used as Training data set and reaming 40%(24,751) as Test data set.

## 5.2 Evaluation Metrics

**Predictive (Classification) Accuracy**:

- This refers to the ability of the model to correctly predict the class label from the training dataset.
- Accuracy = Percentage(%) of testing set examples correctly classified by the classifier

## 5.3 Evaluation Results

The Evaluation Metrics were calculated for using given algorithms and tools below:

- **KNN** –Weka and R
- **Random Forest** – Weka and R

- **ADA Boost** – R
- **XG Boost** – R
- **Ensemble Model** – R and Excel

### 5.3.1 KNN Classifier Results

| K-value | Accuracy – Weka | Accuracy – R |
|---|---|---|
| 9 | 0.7606 | 0.7708 |
| 15 | 0.7599 | 0.7752 |
| 25 | 0.7522 | 0.7616 |

**Table 2 - KNN Accuracy Results**

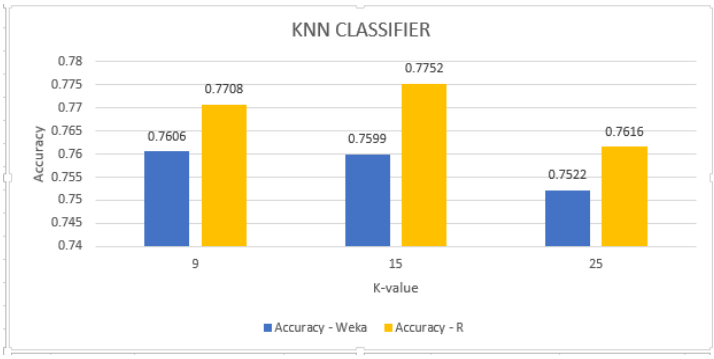


**Figure 4 - KNN Classifier Accuracy – R vs Weka**

### 5.3.2 Random Forest Algorithm Results

| Number of Tress | Accuracy - Weka | Accuracy - R |
|---|---|---|
| 15 | 0.7915 | 0.7738 |
| 25 | 0.7952 | 0.7804 |
| 50 | 0.7977 | 0.7885 |

**Table – 3 Random Forest Accuracy Results**

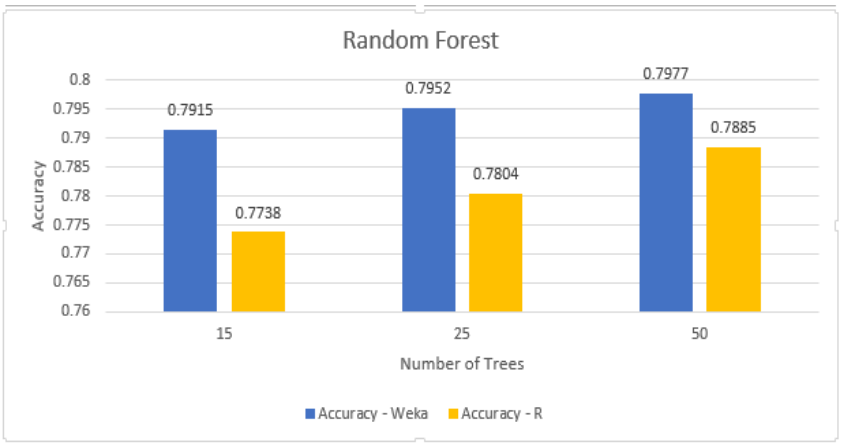


**Figure-5 Random Forest Algorithm Accuracy – R vs Weka**

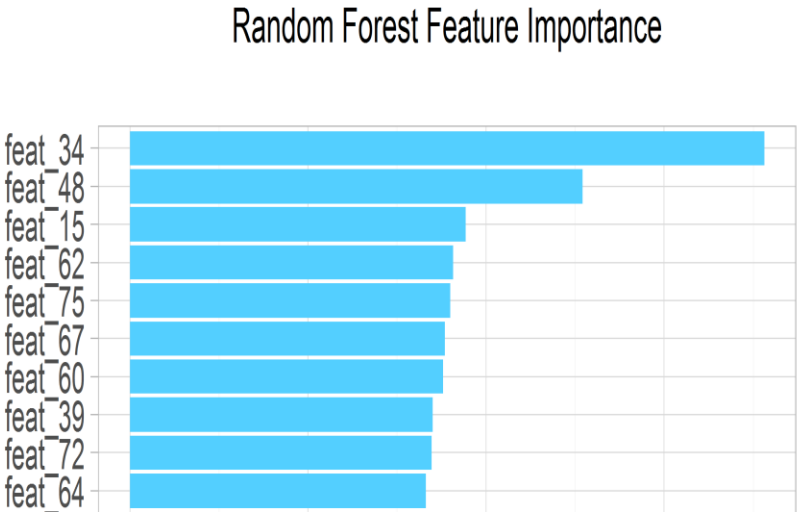


**Figure-6 Top 10 Important Features**

### 5.3.3 ADA Boost Algorithm Results

| Iterations | Accuracy |
|---|---|
| 20 | 0.6219 |
| 40 | 0.6238 |
| 60 | 0.6257 |

**Table – 4 ADA Boost Accuracy Results**

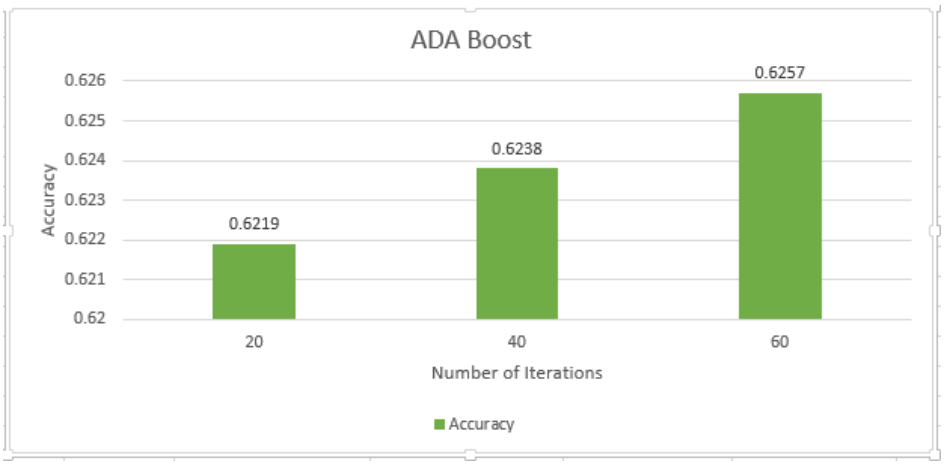


**Figure-8 ADA Boost Accuracy Results**

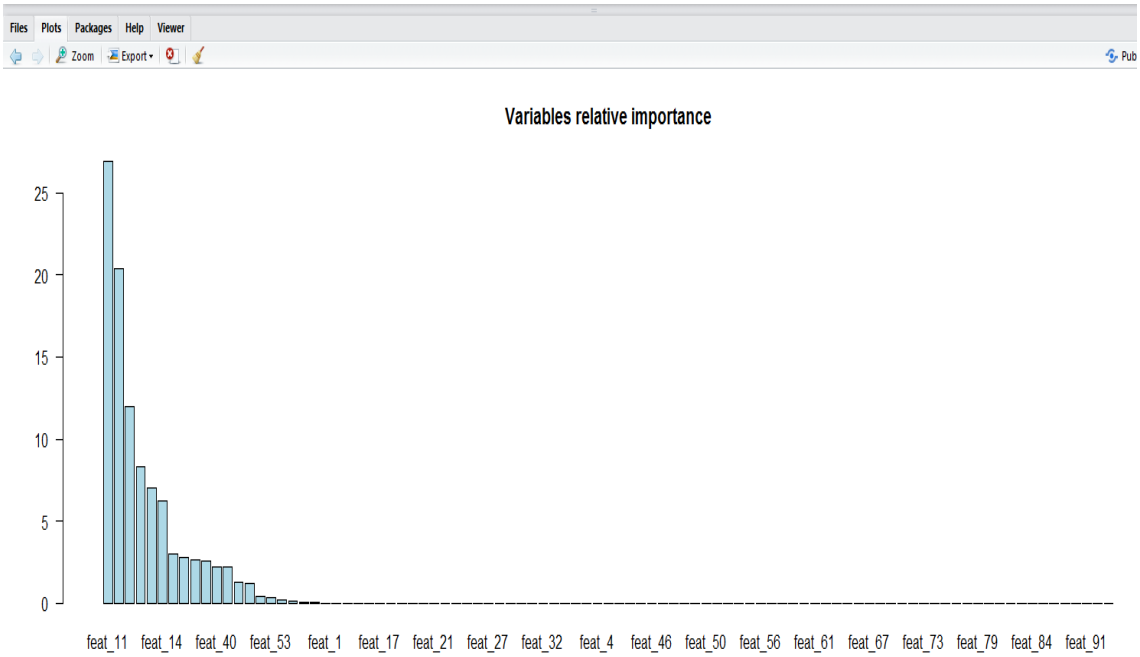


**Figure -7 Top Important Features predicted by ADA Boost**

### 5.3.4 XG Boost Algorithm Results

| Iterations | Accuracy |
|---|---|
| 30 | 0.7651 |
| 50 | 0.7795 |
| 100 | 0.7821 |

**Table – 5 XG Boost Accuracy Results**

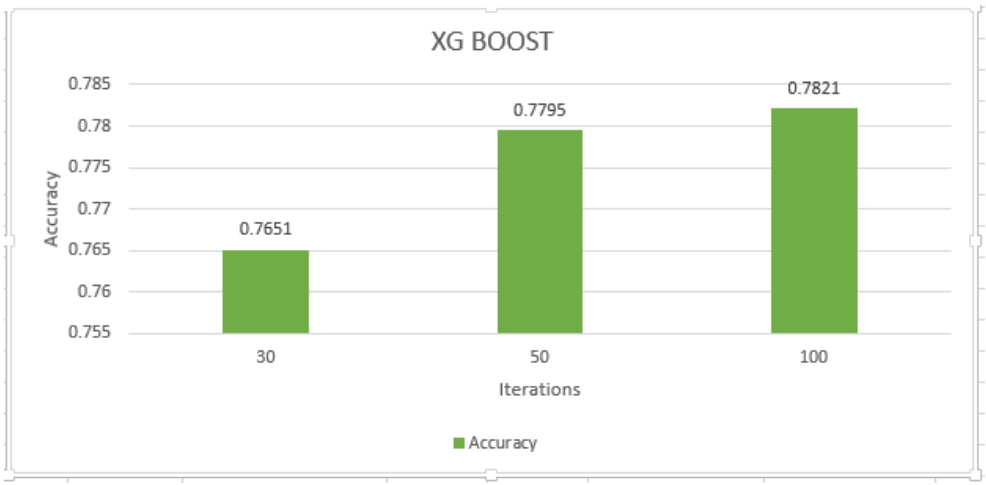


**Figure -9 XG Boost Accuracy Results**

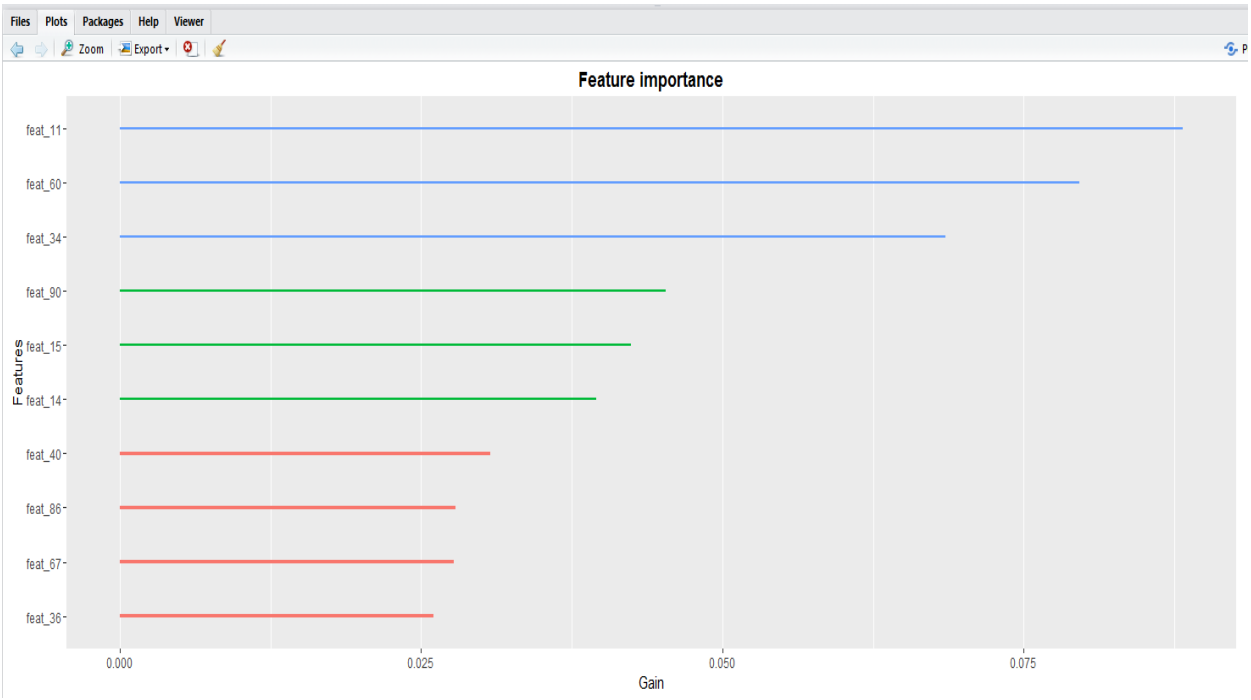


**Figure-10 Top Important Features predicted by XG Boost**

### 5.3.5 Ensemble Model Results

The highest accuracies of each model were used and each algorithm's probabilities are given weight and based on the new probability values the labels are assigned.

**Weighted Average of the Accuracy: [Top XGBoost Accuracy]*0.60 + [Top Random Forest Accuracy] *0.38 + [Top ADA Boost Accuracy] *0.02**

On applying the above accuracy formula, the ensemble model accuracy was **80.7%,** which is a **2.5%** increase with respect to the highest XG Boost accuracy

# 6. Code

The code can be found in below attached document file for all classifiers and ensemble model

Code.docx

# 7. Conclusions and Future Work

## 7.1. Conclusions

- Tree Based Algorithms Random Forest and XG Boost had the best classification accuracy
- Each algorithm predicted its own set of important features from the 93 attributes
- The ensemble model improved accuracy by almost 2%

## 7.2. Limitations

- Unlike other algorithms, KNN classifier does not predict the important features from the dataset
- KNN does not have ability to generate probabilities for each class.

## 7.3. Potential Improvements or Future Work

- In the ADA boost model, additional classifiers can be applied such as decision trees on top of the base classifier to improve performance
- By adding linear regression model onto the ensemble model accuracy can be improved significantly
- With two classes, a good approach is to build BALANCED train and test sets, and train model on a balanced set – randomly select desired number of minority class instances– add equal number of randomly selected majority class

# 8. References


- https://www.ke.tu-darmstadt.de/lehre/arbeiten/studien/2015/Dong_Ying.pdf
- http://www.cse.scu.edu/~mwang2/projects/ML_KaggleCompetition_15s.pdf
- http://blog.kaggle.com/2015/06/15/dont-miss-these-scripts-otto-group-product-classification/