Telstra Network Disruption Prediction

Group Members:

- Vijayalakshmi Vedantham (163050006)
- Vertika Srivastava (163050007)
- Sakshi Maskara (163050041)
- Nidhi Singh (163059004)

Problem Statement

- The goal of our problem was to predict Telstra network's fault severity at a time at a particular location based on the log data available.
- Each row in the main dataset (train.csv, test.csv) represents a location and a time point. They are identified by the "id" column, which is the key "id" used in other data files.
- Fault severity has 3 categories: 0,1,2 (0 meaning no fault, 1 meaning only a few, and 2 meaning many).
- Different types of features are extracted from log files and other sources: event_type.csv, log_feature.csv, resource_type.csv, severity_type.csv.

Motivation

- Study of various data models and their practical implementation.
- Application of Ensembling methods.
- Learn application for practical problem.

Dataset Description

- **train.csv** – the training set for fault severity
- **test.csv** – the test set for fault severity
- **event_type.csv** – event type related to the main dataset
- **log_feature.csv** – features extracted from log files
- **resource_type.csv** – type of resource related to the main dataset
- **severity_type.csv** – severity type of a warning message coming from the log

Data Preprocessing

- Convert features of the form ‘resource type x’, ‘event type x’, ‘location x’, ‘feature x’ etc into categorical variables.

- Create a data frame with the index as unique id.

- Concatenate train and test files.

Feature Enginnering

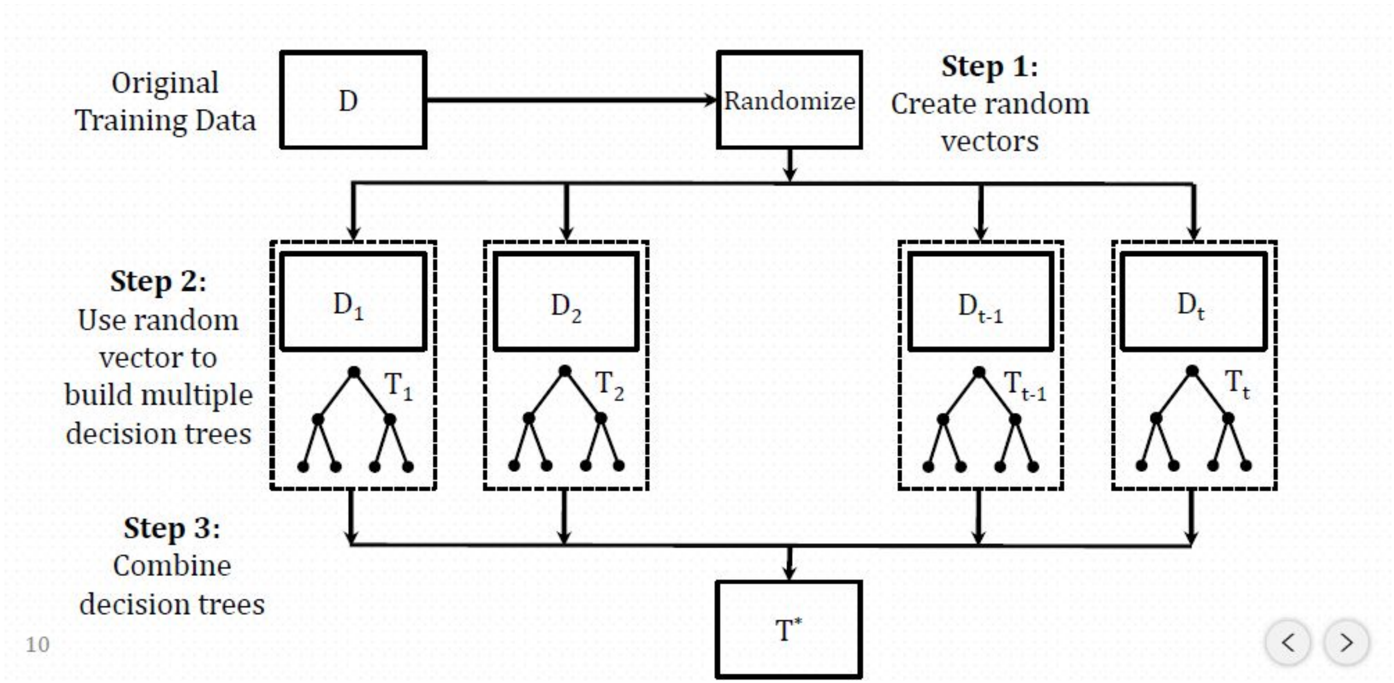
- ✓ Location id based count
- ✓ Number and Reverse number for each location id
- ✓ Normalized number and reverse number
- ✓ Resource type count
- ✓ Frequency based encoding for resource type
- ✓ Event type count
- ✓ Frequency based encoding for event type
- ✓ Log transformed volume
- ✓ Aggregate functions on log volume

Approaches

The models that we applied in our project are:

- Random Forests
- Extra Trees
- XG Boost

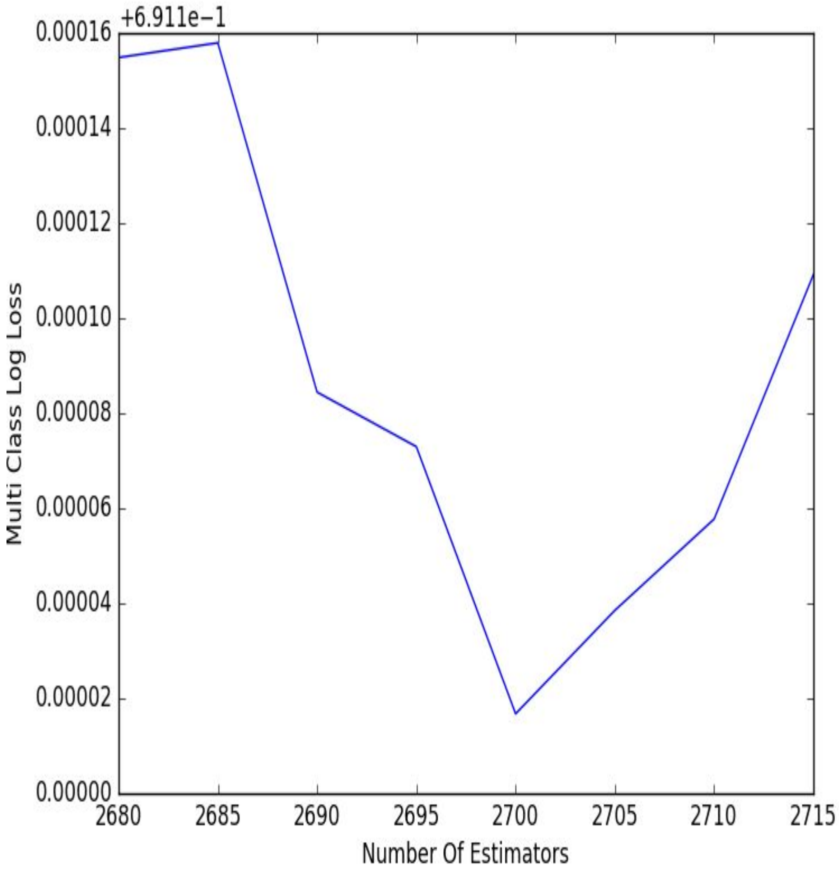
Random Forests



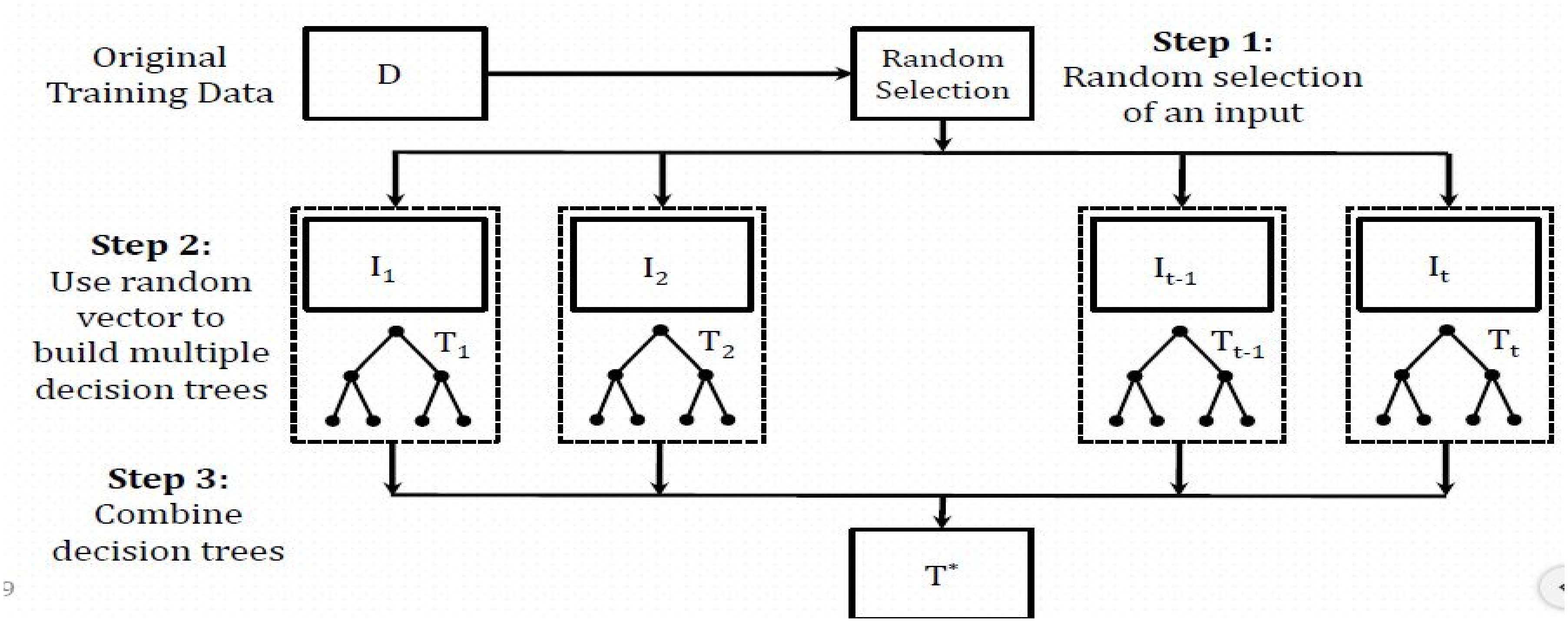
Random forests is a way of combining multiple decision trees which are trained on different parts of training set and different parts of feature set.

Random Forests–Implementation

- We have implemented random forests on our data to predict the severity.
- Different parameters like n_estimators and max_features are varied to get an estimation of parameter value.
- The graph is plotted between No. of estimators i.e. trees and Multi class log loss. The lowest loss is obtained at 2700.



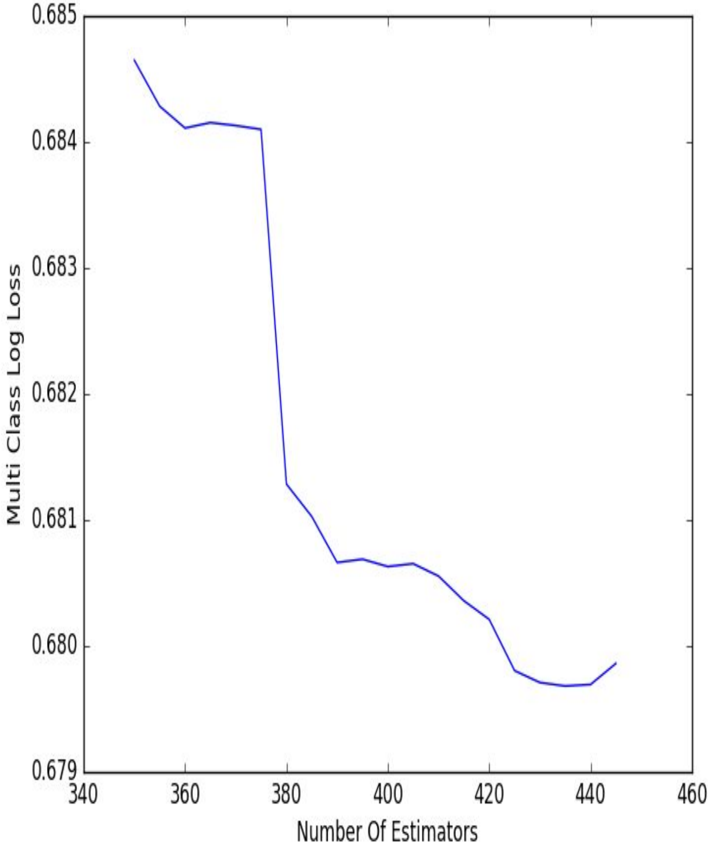
Extra Trees Classifier


Original Training Data
D
Random Selection
Step 1: Random selection of an input
Step 2: Use random vector to build multiple decision trees
I_1
I_2
I_{t-1}
I_t
T_1
T_2
T_{t-1}
T_t
Step 3: Combine decision trees
T^*

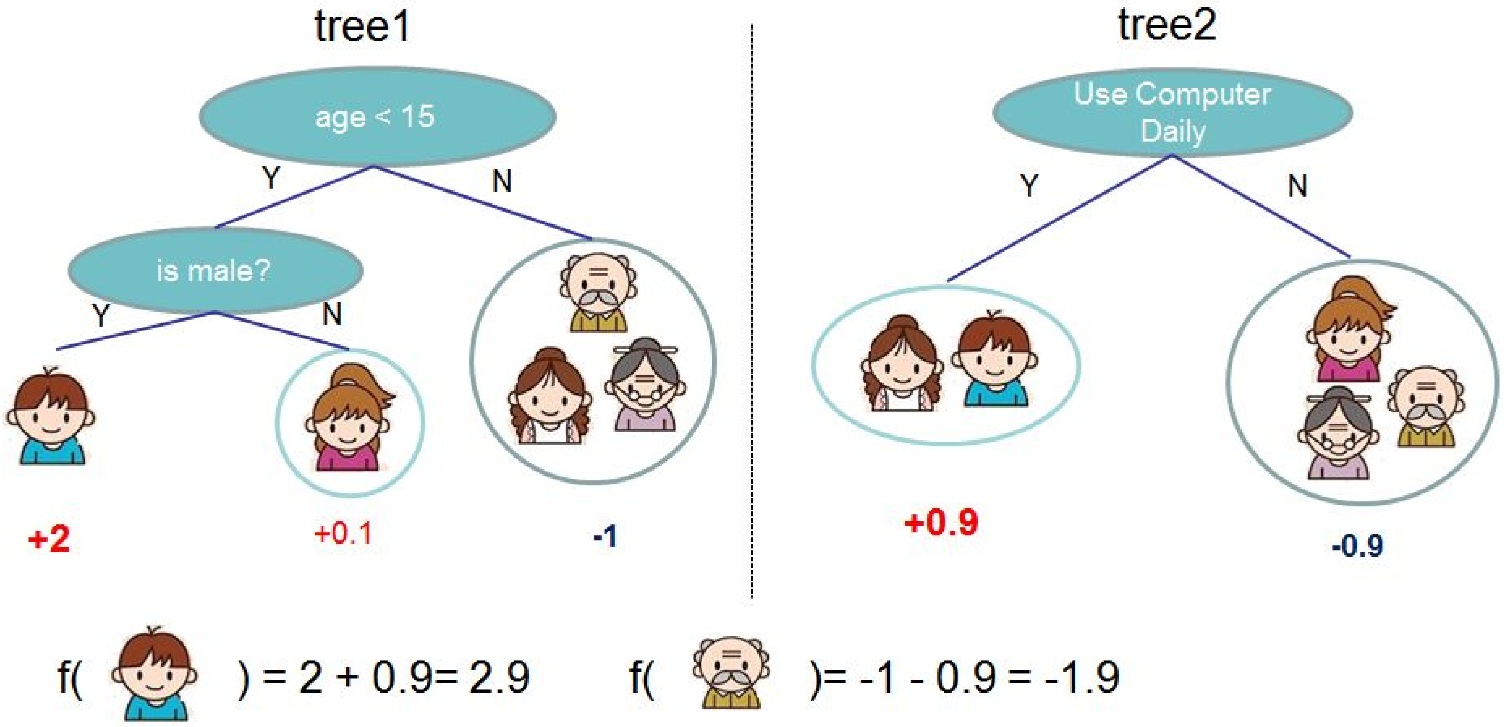
Extra-trees combine classifiers without bagging by using random splits to generate different trees.

Extra Trees Classifier – Implementation

- Here we have implemented extra trees classifier and tried to tune the parameter n_estimators.
- A graph has been plotted by varying no. of estimators to multi class log loss.



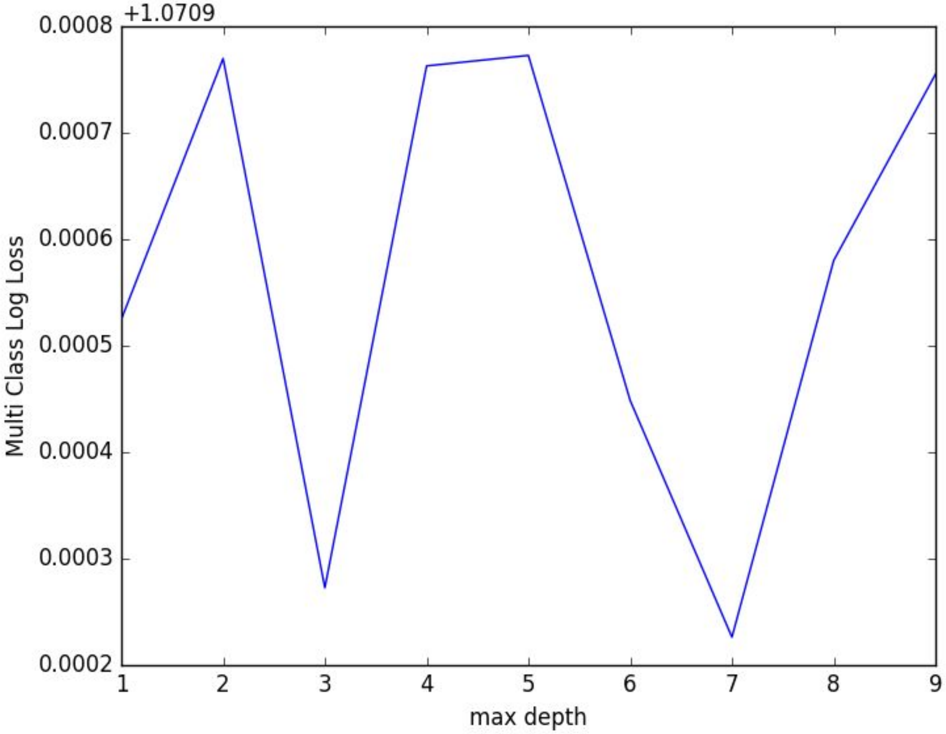
XGBoost Classifier



XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost Classifier – Implementation

- Implemented XGBoost model for data prediction.
- Divided data into 2 parts for train and test.
- Divided train data into 2 parts with probability of 0.8 of taking a row in train and 0.2 of taking a row for watch.
- Used parameters like eta,max_depth,softprop etc.
- Used mlogloss and softprob to get the probability for each class.
- Predicted accuracy on test data.



Observations and Results

Models	Train set accuracy	Test set accuracy
Random Forest	100%	70.08%
Extra Trees	100%	69.89%
XG Boost Classifier	77.8%	73.6%

Challenges

- Feature engineering.
- Combining data from various files into 1 files.
- Tuning models.

Conclusion

- We have implemented different ensembling methods like XGBoost, Random Forest and Extra Trees Classifier to predict the severity of disruption of Telstra Network.
- We have also implemented Neural Networks but it was giving very less accuracy so we dropped it.
- In this project, we have done considerable data preprocessing and feature engineering to model the raw data that was present in multiple data files.

Thank You