CS-GY-9223 - Programming in Big Data Analytics

Comparative Analysis of Machine Learning Models for Classification of Signals Based On Higgs-Boson Experiments

PRESENTERS:
Akshat Tyagi (at3761)
Hamza Mirza (hm1800)
Pulkit Aneja (pa1304)
Vishal Prabhu (vp1179)

New York University

Contents

- Motivation
- Dataset
- Technologies Used
- Data Preprocessing and Cleaning
- Classification Models Used
- Evaluation Metrics
- Confusion Matrix
- Results
- Conclusion
- Challenges Faced and Learning Outcomes
- References

Motivation

- Higgs-Boson is the last and a major piece of the Standard Model of Particle Physics discovered in 2012.
- Explains how particles and forces interact in the universe.
- Could prove theories in physics. For example: Higgs-Boson is responsible for the mass of matter.
- Could change the way we see the universe.
- One step closer to the ultimate goal of proving “The Theory of Everything”.

Dataset

- **Source**: provided by the physicists working on the experiment at CERN.(https://archive.ics.uci.edu/ml/datasets/HIGGS)
- **Task:** Classify whether the event resulted in Higgs-Boson particles or just background noise.
- **Description:** 1 label column, 30 feature columns.
- **Selection:** Features include those measured by detectors and advanced features selected by physicists.

Technologies Used

- Apache Spark (2.1)
    - Spark MLlib
- Tableau
- Apache Zeppelin
- Docker

Data Preprocessing and Cleaning

- **Label Encoding:** Convert the strings in label to double.
- **Removing NA Values:** Drop all na values from the dataset.
- **Normalization and Feature Scaling:** Bring features to a similar scale for easier convergence during optimization process.
- **Model Selection:** Splitting data to train and test data with some random seed.

Classification Models Used

1. Multivariate Logistic Regression with Regularization
2. Decision Tree
3. Random Forest Algorithm
4. Gradient Boosting Tree
5. Multilayer Perceptron (using 3 hidden layers)

Evaluation Metrics

- **True Positive Rate/ Precision:** TP / (TP+FP).
- **False Positive Rate:** FP / (FP + TN)
- **Recall/Sensitivity:** TP/(TP+FN).
- **F1 Score:** 2*((Precision*Recall)/(Precision+Recall))
- **Area Under ROC Curve:** Area under the Receiver Operating Characteristic Curve
- **Area Under PR Curve:** Area under Precision-Recall Curve
- **Accuracy:** Ratio of correct classifications to total number of classifications.

Confusion Matrix

Confusion matrix is the summarization of classification.

PREDICTED

ACTUAL

	P	N
Y	True positives(TP)	False positives(FP)
N	False negatives(FN)	True negatives(TN)

Logistic Regression Confusion Matrix

	Predicted	
Actual	PY	PN
AY	42,258	6,873
AN	12,217	13,643

Value

6,873 42,258

Sum of Value (color) broken down by Predicted vs. Actual.

Decision Tree

Confusion Matrix

	Predicted	
Actual	PY	PN
AY	42,598	6,533
AN	8,051	17,809

Value

6,533 42,598

Sum of Value (color) broken down by Predicted vs. Actual.

Gradient Boosting Confusion Matrix

	Predicted	
Actual	PY	PN
AY	43,731	5,400
AN	7,731	18,129

Value

5,400 43,731

Sum of Value (color) broken down by Predicted vs. Actual.

MLP Confusion Matrix

Actual	Predicted PY	Predicted PN
AY	40,005	9,126
AN	10,513	15,347

Value

9,126 – 40,005

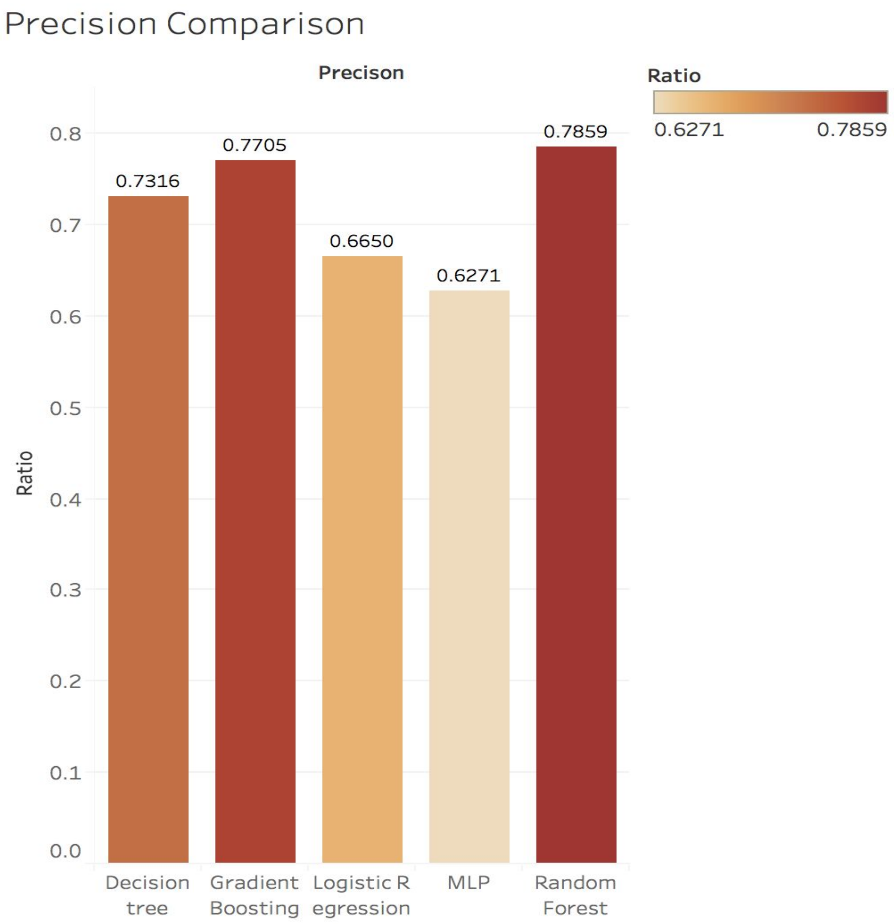
Sum of Value (color) broken down by Predicted vs. Actual.

Random Forest Confusion Matrix

	Predicted	
Actual	PY	PN
AY	44,615	4,516
AN	9,286	16,574

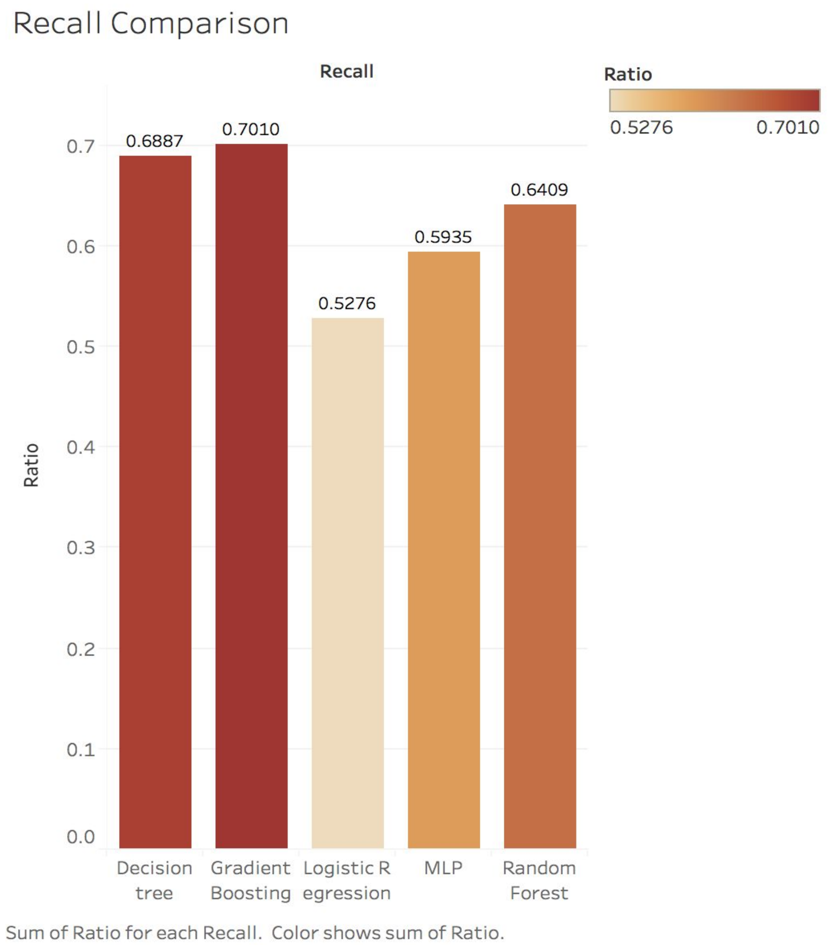
Sum of Value (color) broken down by Predicted vs. Actual.

Precision Comparison

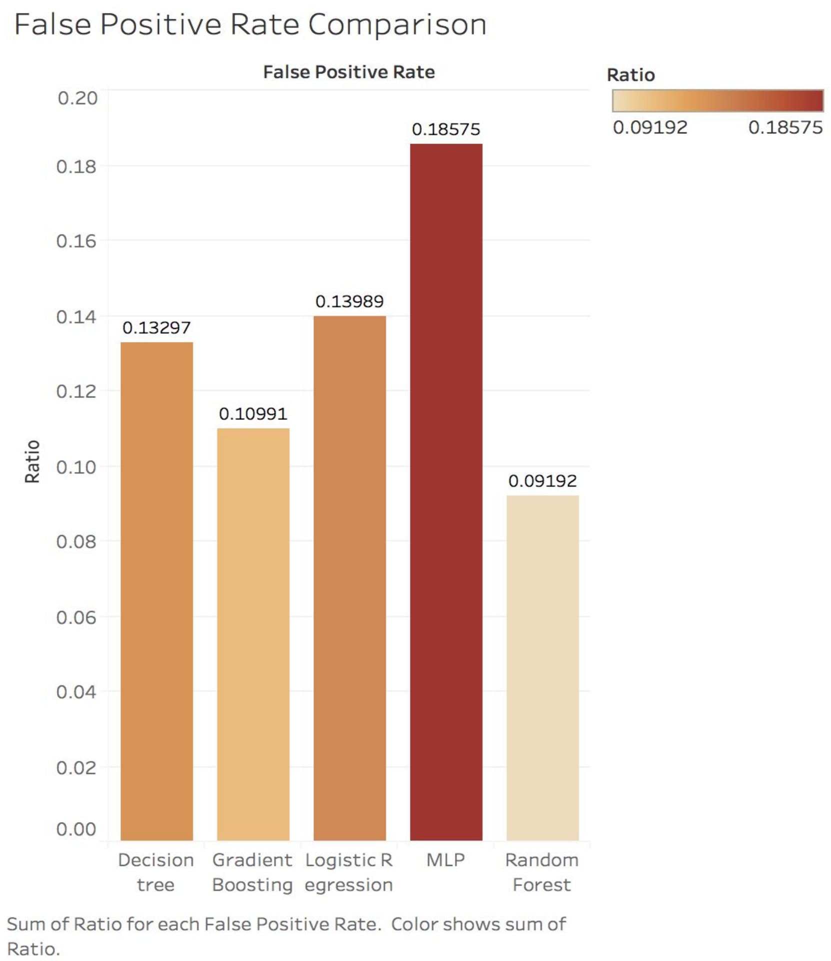


Sum of Ratio for each Precison. Color shows sum of Ratio.

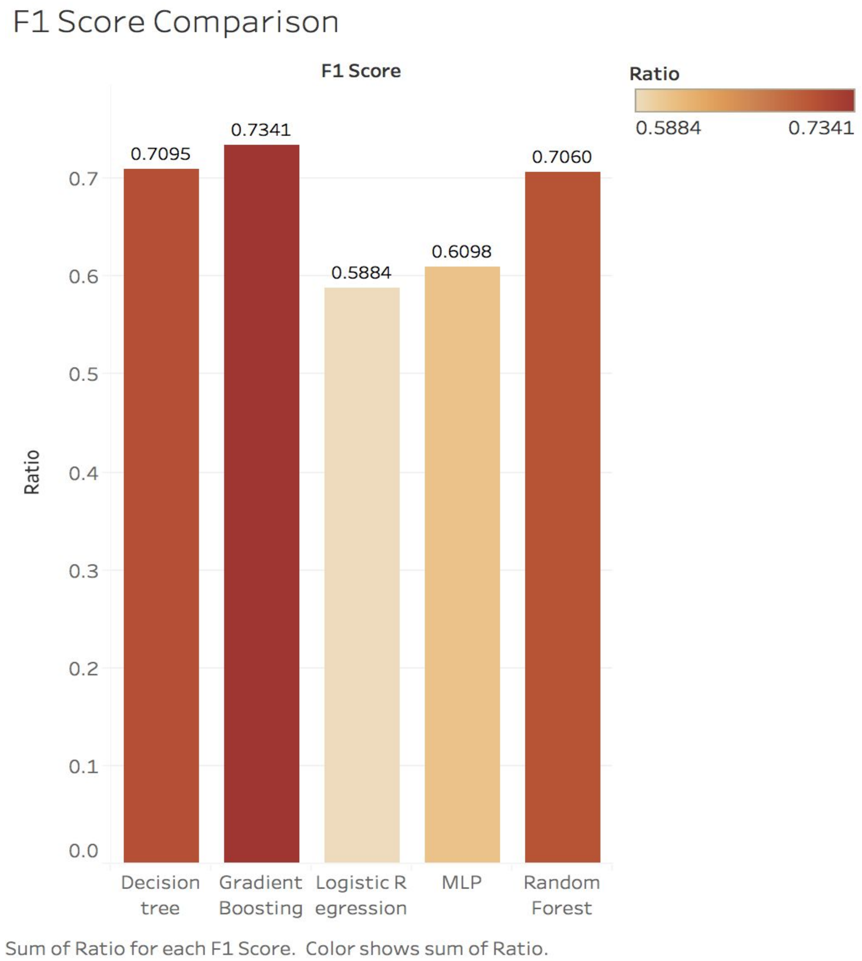
Recall Comparison



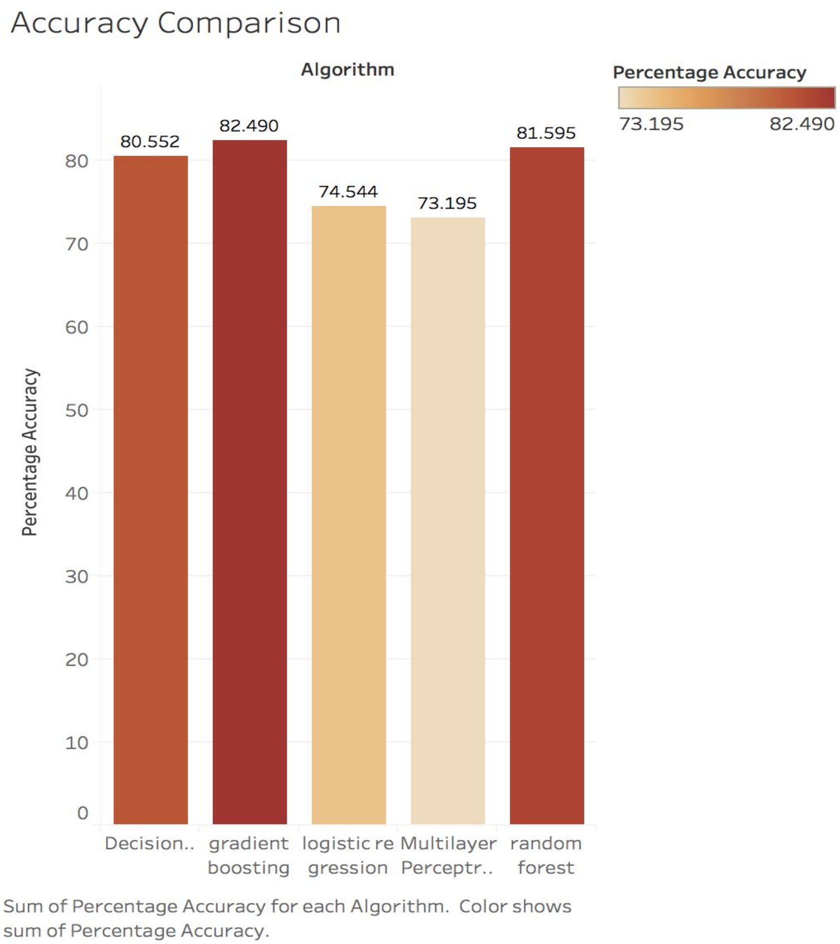
False Positive Rate Comparison



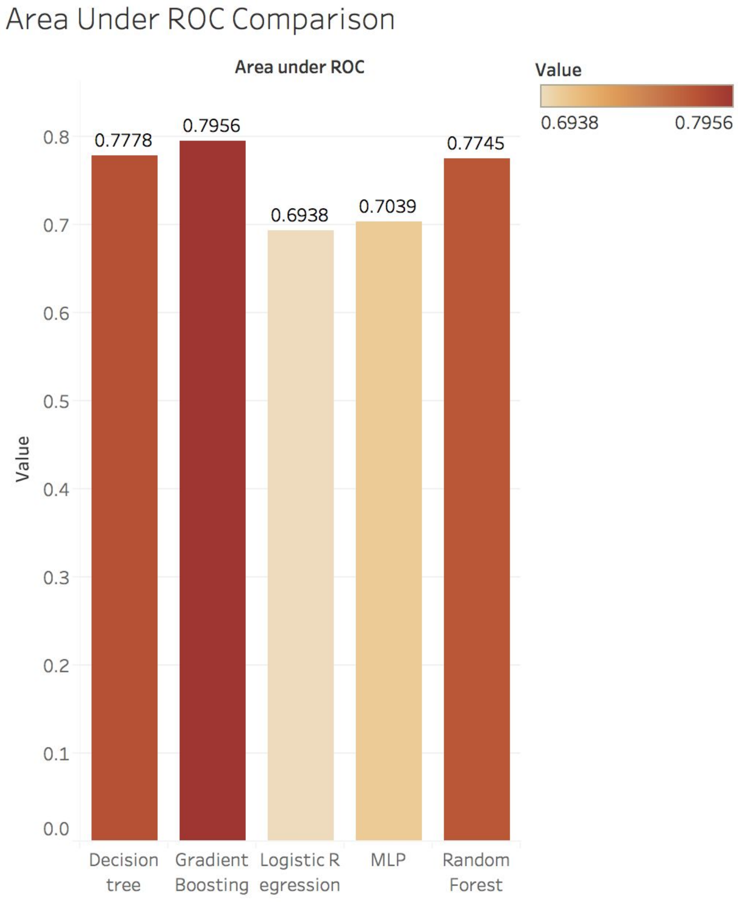
F1 Score Comparison



Accuracy Visualization

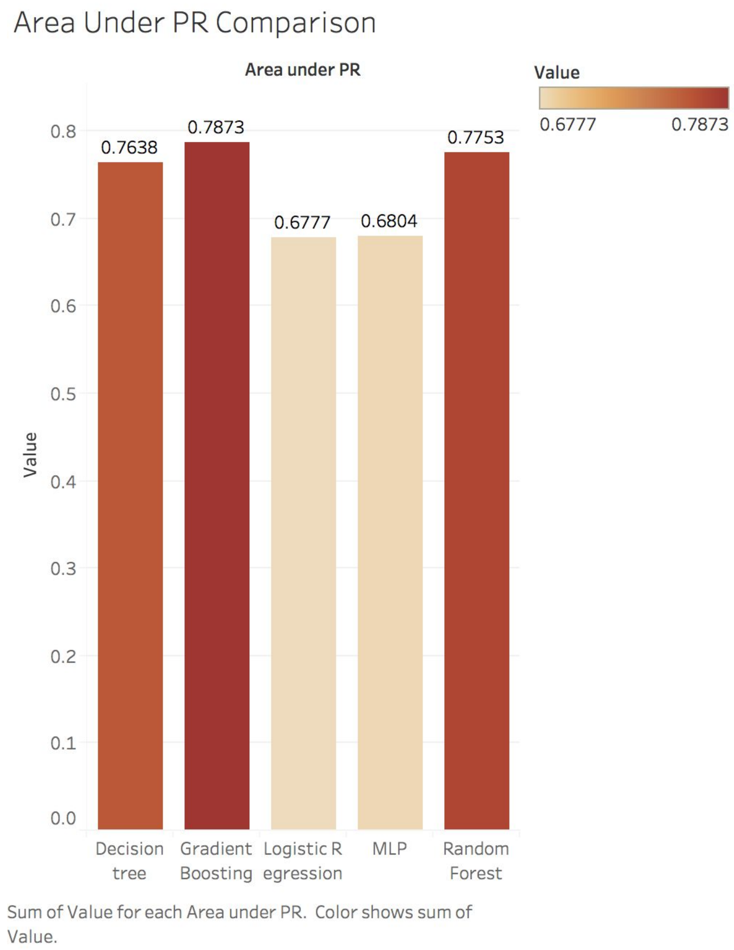


Area Under ROC Curve



Sum of Value for each Area under ROC. Color shows sum of Value.

Area Under PR Curve



Conclusion

- Gradient Boosting Tree performed the best on many metrics.
- Random Forests Algorithm offered better precision and lower FP rate than Gradient Boosting Tree but GBT outperformed in all other metrics.
- Decision Trees had a better F1 Score and Precision than Random Forests.
- Multi-Layer Perceptron had the highest FP rate. Performance could be improved by adding more hidden layers and nodes.
- Slight correlation observed between accuracy, F1 Score, Recall, and TP Rate.
- Randomization affected accuracy.

Challenges Faced

- Launching Apache Zeppelin on NYU HPC.
- Tuning of parameters.

Learning Outcomes

- How to implement ML models on Apache Spark.
- How randomization affects performance.
- Improving performance of algorithms by understanding the data.

References

1. MLlib - Main Guide (https://spark.apache.org/docs/latest/ml-guide.html)
2. Higgs Boson Machine Learning Challenge (https://www.kaggle.com/c/higgs-boson)
3. UCIML - Higgs (https://archive.ics.uci.edu/ml/datasets/HIGGS).