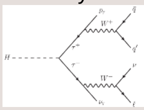
HIGGS BOSON CHALLENGE

"The Gradients"-Abjasree S, Poojan Smart, Niranjan Solanki, Vaishnav Panuganti.

Motivation and About the Project

Motivation:

Higgs boson is the particle which is proposed to give mass to other elementary particles.
This discovery contributes to our understanding of the origin of mass to the subatomic particles.
The Higgs boson is discovered from its decay channels after it is produced in the proton-proton collisions.
The decay channel of the Higgs boson through tau-tau particle is considered as the signal.



About the Project:

The goal was to optimize the analysis.
The problem is that given the predictors classify it as signal or background.
To improve the signal selection, objective here it to improve Recall /TPR along with the reducing FPR.
The objective they have formally provided in the problem is to maximize the AMS where s is TPR and b is FPR

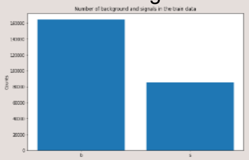
$$\text{AMS} = \sqrt{2\left((s+b+b_{\text{reg}})\ln\left(1+\frac{s}{b+b_{\text{reg}}}\right)-s\right)}$$

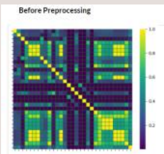
Data and Labels

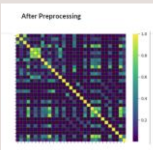
Data: The data is the simulated data from the ATLAS experiment at CERN

Predictors: There are 30 features which is used to fit the model

Labels: The classification problem here is to predict as "signal" and "background"







The data here is an imbalanced data
The data set contains so many undefined values i.e., -999
We used median imputation so that the correlation due to undefined values can be eliminated.
We also used Standardscaler() to scale the data points

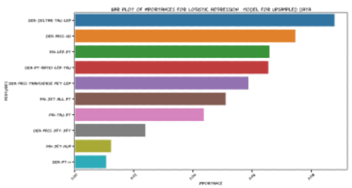
Models and Results

Approaches:

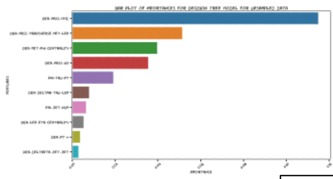
- The aim was to improve the recall or TPR so that the signal is predicted accurately
- We took two approaches to tune the hyperparameters using the scoring as recall and f1 score

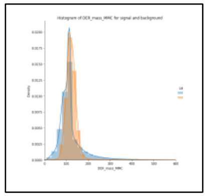
Base Models Used

Logistic regression:



Decision trees:





- These the 10 most important features we got for the base models
- The feature importance is calculated from the permutation importance

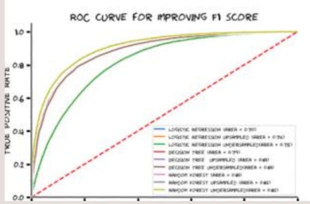
Ensemble Models Used

- We tried 3 ensemble model i.e., bagging, gradient boosting and random forest
- The random forest was giving the best AMS values
- We tuned the hyperparameters for random forest with scoring as recall and f1 score
- The feature importance is same as that of decision tree for the ensemble models

Model	Recall	F1 score	AUC score	AMS
Logistic Regression - No Imbalance Correction	0.533	0.595	0.698	0.874
Logistic Regression - Upsampling	0.764	0.667	0.744	0.887
Logistic Regression - Downsampling	0.77	0.668	0.745	0.8901
Decision Tree -No Imbalance Correction	0.703	0.732	0.794	1.157
Decision Tree - Upsampling	0.8	0.743	0.807	1.034
Decision Tree - Downsampling	0.807	0.741	0.805	1.018
Bagging - No imbalance Correction	0.709	0.744	0.803	1.232
Bagging - Upsampling	0.815	0.753	0.816	1.066
Bagging - Downsampling	0.821	0.756	0.818	1.073
Gradient Boosting - No imbalance correction	0.71	0.747	0.805	1.235
Gradient Boosting - Upsampling	0.807	0.757	0.819	1.088
Gradient Boosting - Downsampling	0.802	0.758	0.82	1.077
Random Forest - No imbalance correction	0.709	0.751	0.807	1.292
Random Forest - Upsampling	0.805	0.766	0.824	1.14
Random Forest - Downsampling	0.805	0.766	0.824	1.14

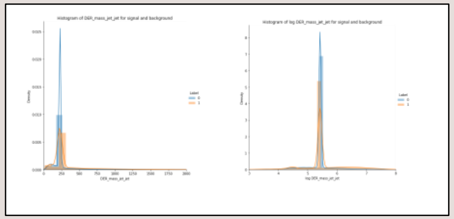
Conclusions

- ✓ Up/Down sampling gave better accuracy, recall, f1 but AMS went down for decision tree-based models.
- ✓ For Logistic regression it slightly improved AMS.
- ✓ Tuning hyperparameters using F1 score did not provide any significant benefit.
- ✓ **Feature Importance**: Derived columns seems to be more important compared to Primary columns.



Future Work

- ✓ Feature Engineering: Log transform for long tailed distributions (14 features) so it will be more symmetric.



- ✓ Better estimation for AMS (Bootstrapping for maximising AMS for different ensemble models)
- ✓ Better way to incorporate invalid values so that its importance is not lost

References

1. documentation_v1.8.pdf (in2p3.fr)
2. G. Aad et al., Phys.Lett., vol. B716, pp. 1–29, 2012
3. The ATLAS Collaboration, Tech.Rep, ATLAS-CONF-2013-108, November 2013
4. CERN