Dataverse Hack

Hackathon Final Submission

Brief about the Approach

1. The Data was first explored to see the data types and structure.
2. The problem was identified to be supervised and of Classification type.
3. Many models were tried before the Final Submission. XGBoost model gave the best results.
4. XGBoost Classifier was used to classify the Target Variables.
5. RandomizedSearchCV was used to tune the parameters to get the best result.

Data pre-processing

1. The Policy ID was first removed in both the Train data and the test data.
2. ‘max_torque’ and ‘max_power’ columns had values in object type. .apply(), .split() and lambda were used to get the required values from the columns.
3. Later the values were changed to numeric data type using pandas.to_numeric() function.
4. LabelEncoder from sklearn was used to convert the categorical data into numeric data. .select_dtype() was used to identify the object type data columns.
5. The data had disproportionate Target Labels. Hence the data sampled using .sample() function.
6. The Features/attributes were later separated from the Target Variables.
7. Heatmap was plotted to check correlation between the attributes.
8. Train_test_split() was used to split the Training Data into Training and Validating Set.
9. The Policy ID column was again loaded for submission purpose.
10. The model was built using XGBoost and RandomizedSearchCV.
11. Submission csv was obtained using pandas.to_csv() function.

Final Model

1. The final model is XGBoost.
2. Various parameters were tested using RandomizedSearchCV to obtain the final parameters for the Final Model. ‘f1’ was used as scoring parameter.

```
XGBClassifier(colsample_bytree=0.8, gamma=2,
min_child_weight=10, n_estimators=600, nthread=1,
subsample=1.0)
```

The above estimator was deemed best by RandomizedSearchCV.

3. best_estimator_ and best_score_ were used to get the score and best estimator.