FE Technique 1

* **Filter Method** is applied using 3 techniques

1. Removal of constants, quasi-constants, and duplicate features: – Threshold variance is set to 1% which filters 7 features as they don’t contain much information.
2. Removal of Correlated features:– as per high correlation rules 4 features are filtered as they constitute similar possibilities.
3. Using Statistical methods

Overview

**Objective:** The signature detection system identifies patterns of traffic or application data presumed to be malicious. The objective is to analyze the dataset, investigate and evaluate the results, predict the overall performance and accuracy of the model.

EDA – Feature-wise Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature - Numerical - No Outliers** | **Feature - Numerical - Outliers** | **Feature - Categorical** | **Ignored?** |
| duration | hot | protocol\_type | No |
| src\_bytes/dst\_bytes | num\_compromised | service | No |
| same\_srv\_rate | num\_root | flag | No |
| dst\_host\_count | num\_file\_creations | land | Yes |
| dst\_host\_srv\_count | count | wrong\_fragment | No |
| dst\_host\_same\_srv\_rate | srv\_count | urgent | Yes |
| dst\_host\_rerror\_rate | serror\_rate | num\_failed\_logins | No |
| dst\_host\_srv\_rerror\_rate | srv\_serror\_rate | logged\_in | No |
|  | rerror\_rate | root\_shell | No |
|  | srv\_rerror\_rate | su\_attempted | Yes |
|  | diff\_srv\_rate | num\_shells | Yes |
|  | srv\_diff\_host\_rate | num\_access\_files | No |
|  | dst\_host\_diff\_srv\_rate | is\_guest\_login | No |
|  | dst\_host\_same\_srv\_port\_rate |  |  |
|  | dst\_host\_srv\_diff\_host\_rate |  |  |
|  | dst\_host\_serror\_rate |  |  |
|  | dst\_host\_srv\_serror\_rate |  |  |

Dataset Understanding

* **How many features:** The given dataset contains 40 features which includes CLASS feature as the target attribute having symmetric categorical and doesn’t contain missing values.
* **Size of the dataset:** The total number of records are 22544 which includes protocol\_type (3 unique), service(64 unique) and flag (11 unique) being the categorical features in the dataset.
* **Missing data and Preprocessing:** The attributes ‘count’ (#18) and ‘srv\_count’ (#13) have missing values are replaced with mean of the numerical attributes by target CLASS.
* **Outliers and incompatible attributes:** Considered attributes as outliers if they are behind or beyond – 3x -ve/+ve SD or 1.5x IQR. (Attribute level outliers are explained detailed in EDA portion)

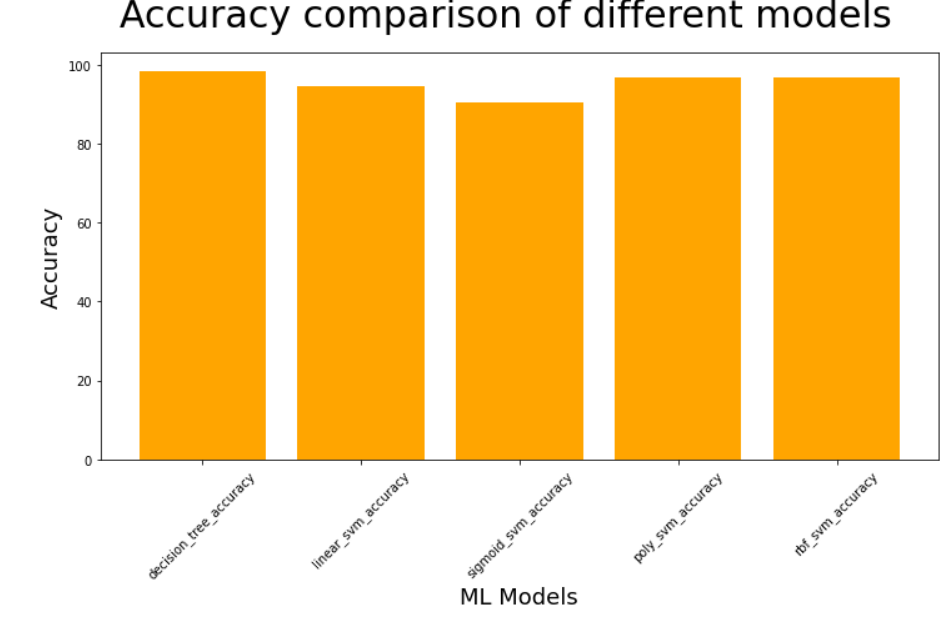
Improvements

There are still many things we can try, which are listed as follows:

* Do more exploratory data analysis to observe the feature interactions.
* Add in categorical features to be the predictors.
* Implement advanced feature selection pipeline to select relevant features.
* Try to use different classifiers to do the model comparison.

Comparison ML with FE

The below comparison is between decision tree and various SVMs – Linear, Sigmoid, Poly, RBF.

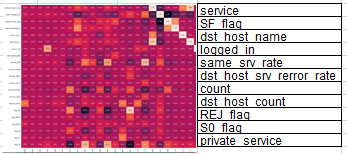


Conclusion

We have applied and compared various ML workflow along with exploratory data analysis to final model evaluation. The major feature - src\_bytes contribute maximum towards the DT model

FE Technique 2

**Top 11 identified features based on the heatmap Correlation with the target variable** –



Decision Tree

1. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions.
2. Split the dataset across training (60%) and testing (40%) datasets
3. Decision Tree Model Accuracy: 98.33666001330673
4. Decision Tree Cross Validation Accuracy: 98.32 %

SVM

1. Support Vector Machine is a supervised machine learning algorithm that can be used for both classification/regression challenges.
2. It is more productive in high dimensional spaces.
3. SVM RBF Model Accuracy: 96.66223109336882