

"THE OUTLIERS" PRESENTS

Anomaly Detection Challenge -3 Malware Classification

**NITHISH
RAGHUNANDANAN
VISHAL BHALLA**

Agenda

- ▶ Introduction
- ▶ System Pipeline
- ▶ Data Preprocessing & Analysis
- ▶ Feature Engineering
- ▶ Model - Tuning & Evaluation
- ▶ Kaggle Results
- ▶ Conclusion
- ▶ Key Takeaways

Introduction

- ▶ Aim: Classify malware into predefined families (0...9) or outlier (10).
- ▶ Problem Type: Multi-class Classification
- ▶ Samples
 - ▶ Training Set: 2056 (1627 after Removal of Missing Training Data)
 - ▶ Test Set: 1676
- ▶ Number of Features: PE Info(Static Analysis) and VT Info(Antivirus Signatures)
- ▶ Classification
 - ▶ There are 11 decision classes: 0 to 10.

System Pipeline



- ▶ Data Visualization
 - Tabulate & analyse the structure of the data.
- ▶ Feature Selection
 - Encoding selected parts of the data as features.
- ▶ Evaluation of Models
 - Classification accuracy using K Fold Stratified Cross Validation.

Data Preprocessing

- ▶ Removal of Missing Training Samples
 - ✓ No Static Analysis and Anti Virus Signatures for 429 samples of the training set
 - ✓ Removed the training samples from the training phase

Data Analysis

- PE Info
 - Multiple pe_section, pe_resources, pe_import, pe_timestamp, rich_header sections per sample.
 - [Sample File](#)
 - Analysis of Sections
 - [Analysis](#)
- VT Info
 - Single Signature for some of the 75 Anti Viruses per sample.
 - [Sample File](#)
 - Frequency Analysis for Anti Virus Signatures.
 - [Analysis](#)

Feature Engineering

- ▶ VT Info
 - ✓ Selecting sub set of Anti Viruses with high variance.
 - ✓ Selecting all Anti Viruses.
 - ✓ Reduction of Null Feature Vectors.

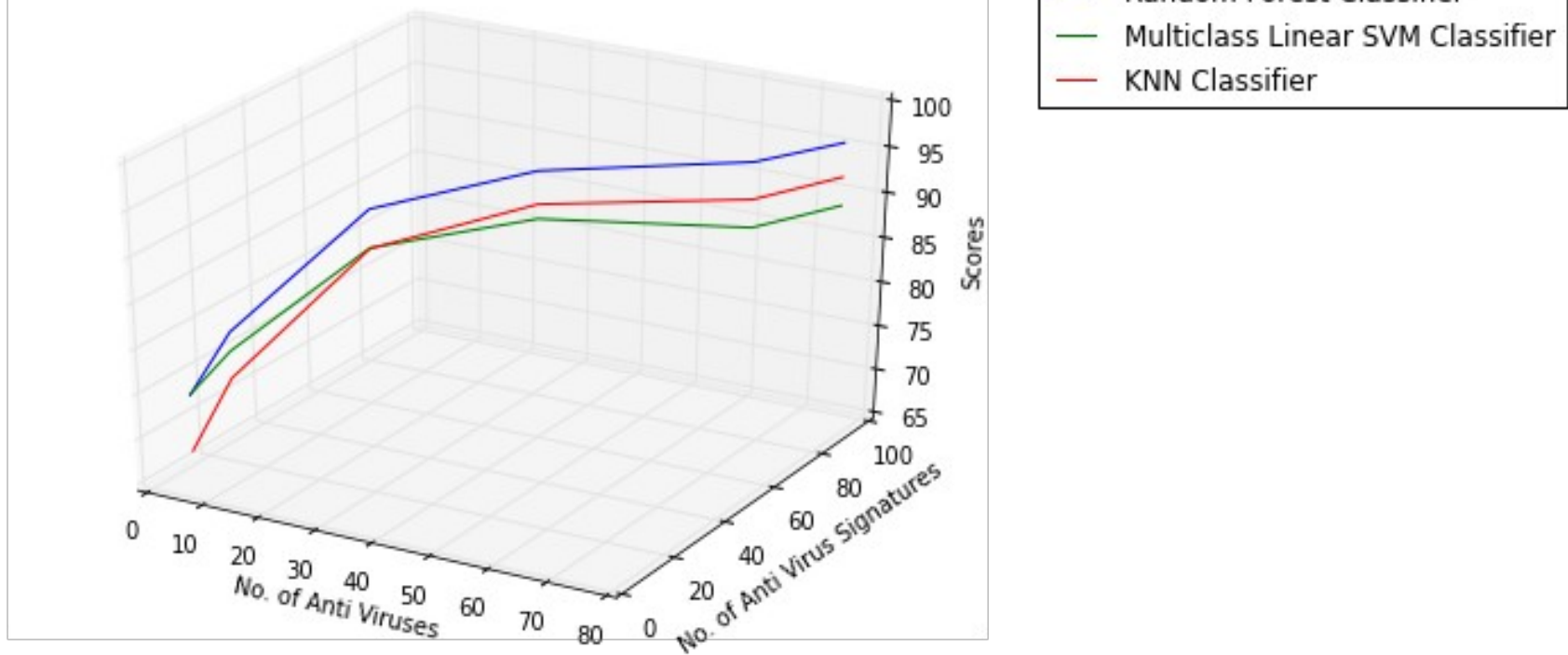
- ▶ Variance of Anti Virus
 - ✓ $\text{Variance} = \text{Total No. of Samples} / \text{No. of Unique Signatures}$

Feature Engineering(2)

- ▶ Encoding of Features
 - ✓ One Hot Encoding
 - Blow up of features.
 - 75 anti viruses * 203 Signatures on average
 - ✓ Binary Encoding
 - Presence/Absence of Anti virus signature for the sample
 - ✓ Frequency Encoding
 - Variance of Anti virus/ Total Occurences for the sample

Feature Engineering(3)

Analysis of Number & Signature of Anti Viruses



Feature Engineering(4)

► PE Info

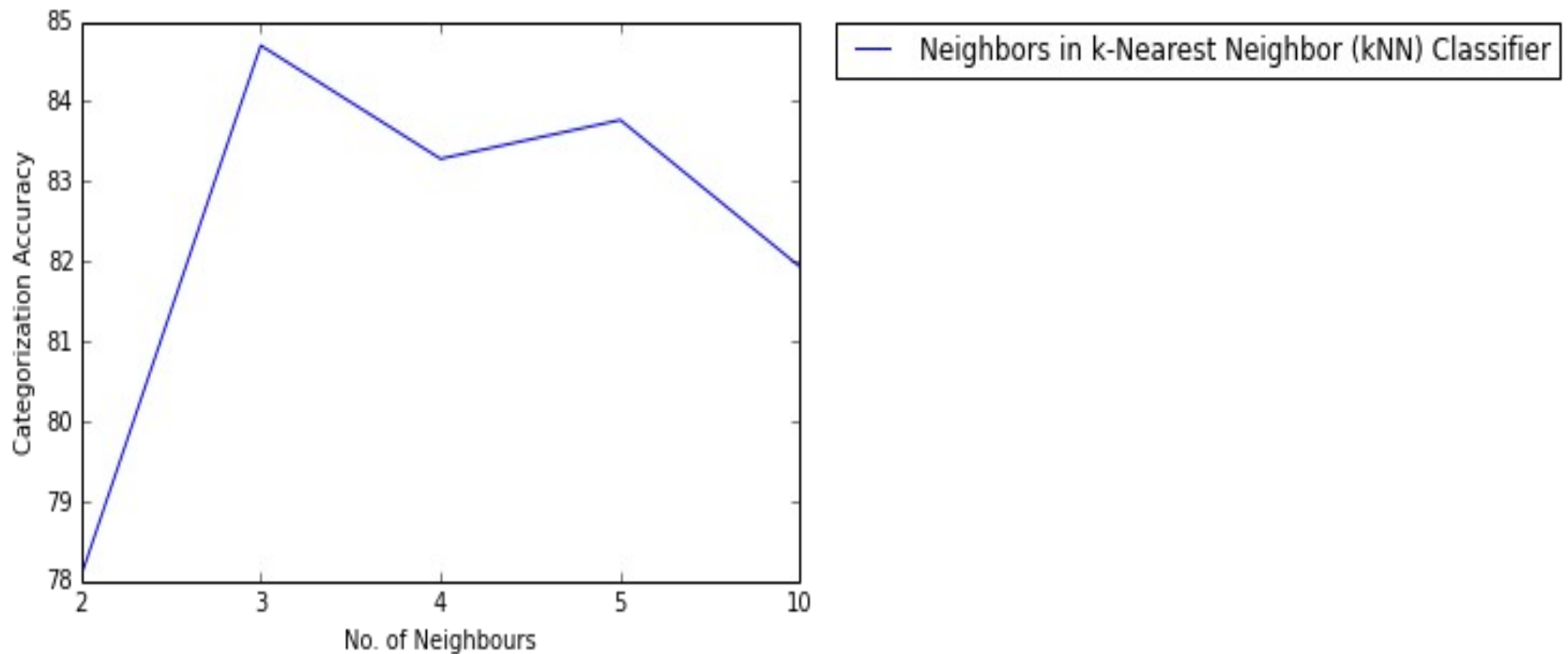
- ✓ Most of the sections are repeated per sample.
- ✓ Mean, Median & Mode could be tried on Numerical Features.
- ✓ Mean & Median gave best results.
- ✓ Increasing number of features did not necessarily give better results compared to pure VT Info Encoding.
- ✓ Selected just Entropy.

Model

- ▶ We tried different types of Multi Class Classification Models to fit our data
 - ▶ Random Forests
 - ▶ Multi Class SVM with Linear Kernel
 - ▶ K Nearest Neighbors with 3 Neighbors

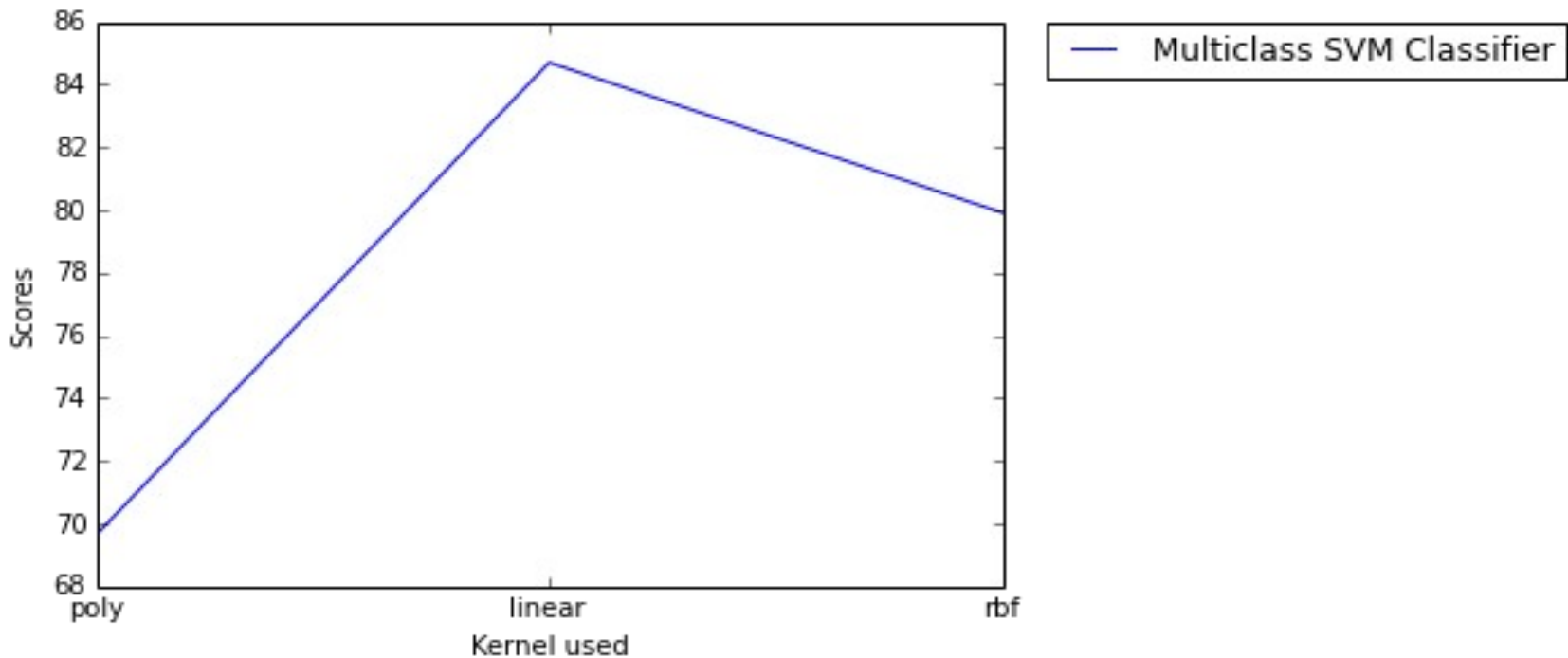
Tuning Model Parameters

K Nearest Neighbors (kNN)



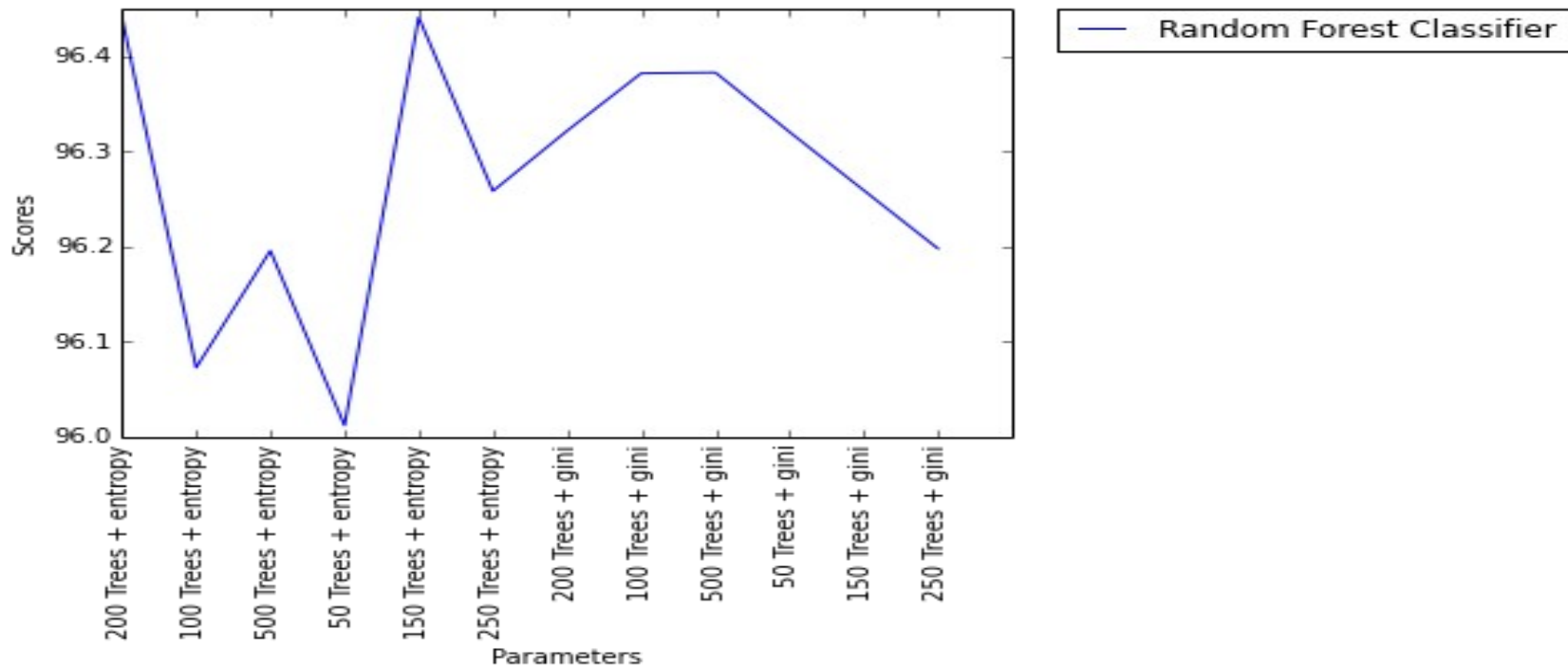
Tuning Model Parameters(2)

Multi Class Support Vector Machines (SVM)



Tuning Model Parameters(3)

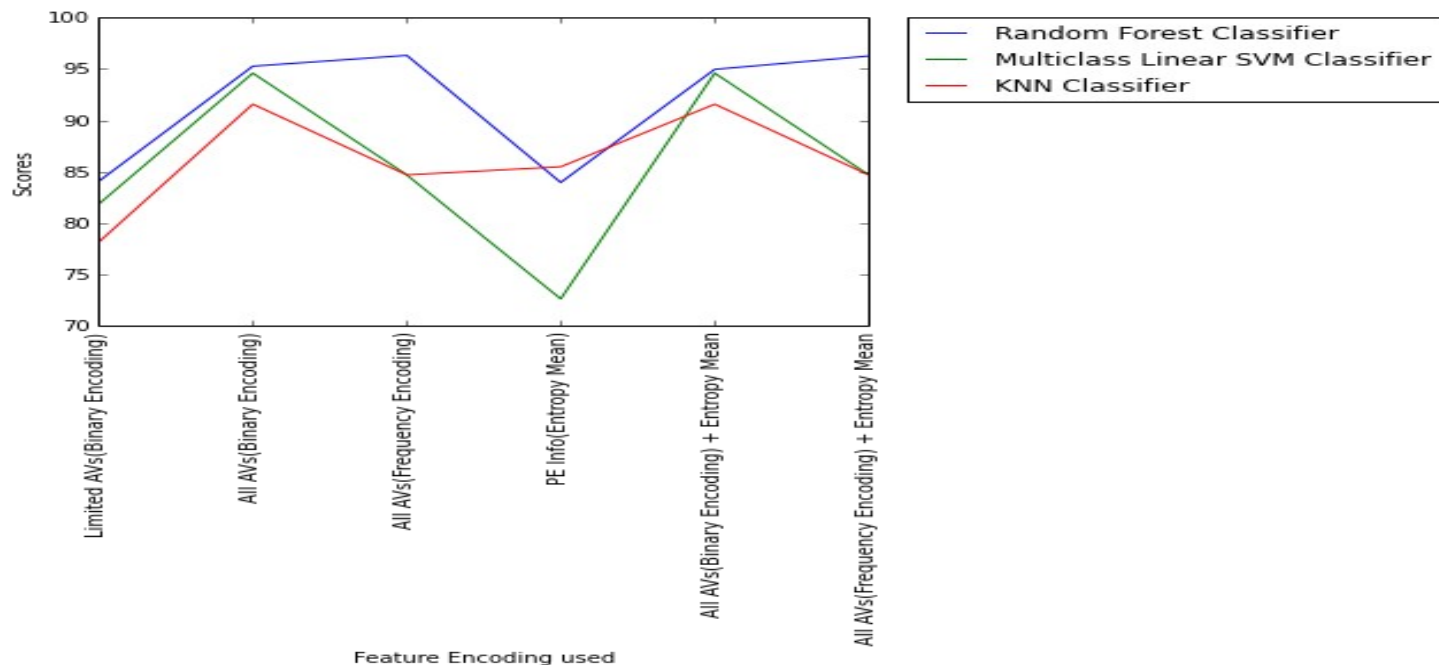
Random Forests (RF)



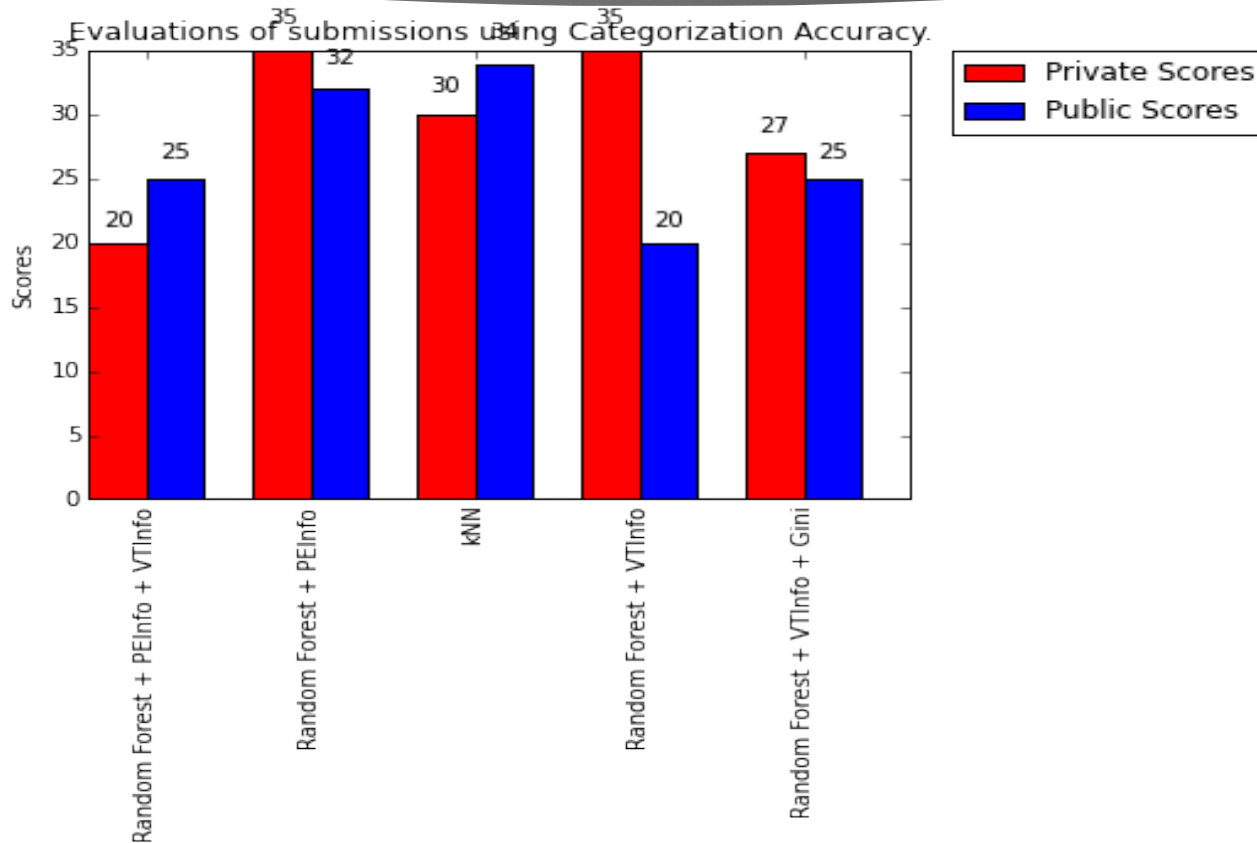
Evaluation of the Models

Criteria

Classification Accuracy on Stratified K-Fold Cross Validation with a split of 4:1.



Kaggle Results



Conclusion

- ▶ The best results were observed for Random Forest (100 trees with Gini as a criterion) on Frequency Encoded all Anti Viruses' VT Info data.
- ▶ The addition of Entropy Mean/Median from PE Info gives almost the same result as pure Frequency based VT Info encoding on all Anti Viruses.
- ▶ KNN & Multi Class SVM worked better on Binary Encoding of VT Info.

Key Takeaways

- ▶ Visualization for analysis of data set for feature engineering.
- ▶ Inherent structure of our data consisted of mainly categorical features. Random Forest Classifier works well on categorical features.
- ▶ Adding more features could result in overfitting.



Questions?

Thank You !