Microsoft Malware Prediction

Springboard Data Science Career Track
Capstone Project 2

Marco Riva
August 6, 2021

Motivation

Malware is a serious threat to billions of customers privacy and security

Can we use Machine Learning to Prevent Malware infection?

- Predictive tools for malware detection are key to preserving customer privacy!

Problem Formulation

- Supervised Learning
- Binary Classification: Malware vs No Infection
- Balanced dataset

Objectives

- Utilize historical computer data provided by Microsoft to predict if a Windows machine will be infected by malware
- Identify features correlated with Malware infection and provide actionable insights

Evaluation Metric

- Receiver Operating Characteristic Area Under the Curve (ROC AUC)

Data Source

- https://www.kaggle.com/c/microsoft-malware-prediction/data

This presentation

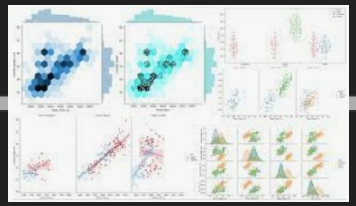
Data Acquisition

CSV
9M+ rows
83 columns

Data Wrangling

Outliers
Errors
Inconsistencies
High Cardinality

Exploratory Data Analysis



Cleaned Data Exploration

Feature Engineering

Processing
Encodings
Transformers

Modeling

Binary Classification

Insights

What we've learned

Data Wrangling

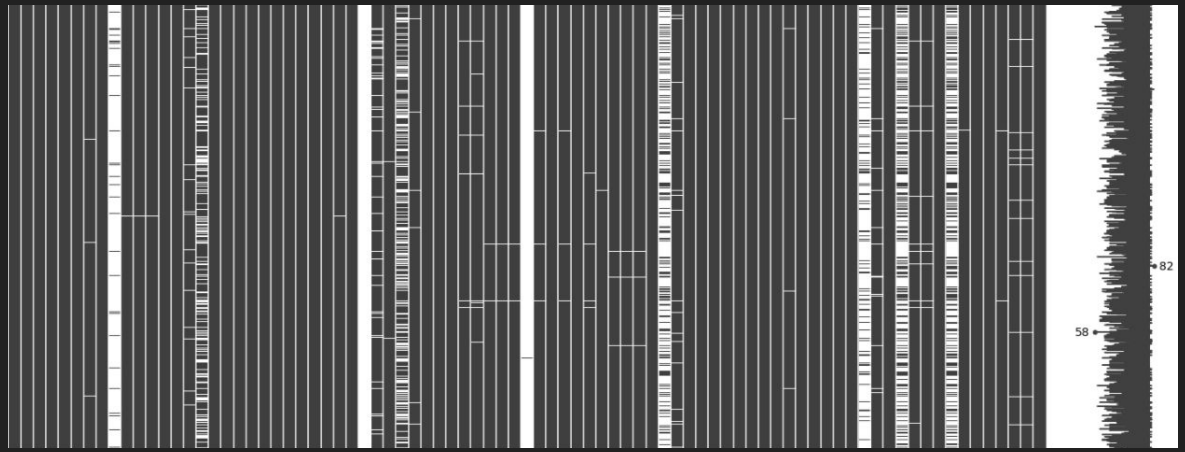
What Features?

- Machine Specs and Operative System (OS)
- Microsoft Defender Specs
- AntiVirus & Security Settings
- Browser & Apps
- Geographical Information
- Device Census File Data

What Target?

- Balanced binary variable
- Positive class implies malware presence

Data Wrangling - Missing Values

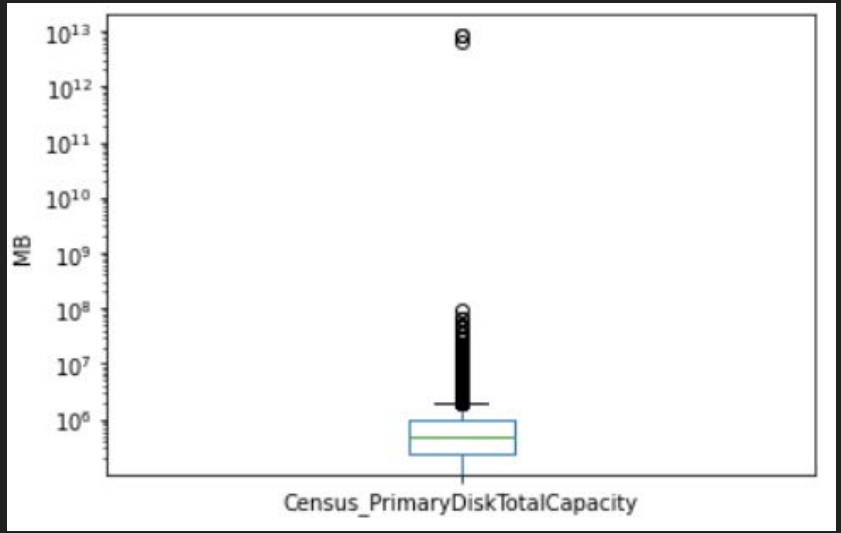


	missing_count	%
PuaMode	8919174	99.97
Census_ProcessorClass	8884852	99.59
DefaultBrowsersIdentifier	8488045	95.14
Census_IsFlightingInternal	7408759	83.04
Census_InternalBatteryType	6338429	71.05
Census_ThresholdOptIn	5667325	63.52
Census_IsWIMBootEnabled	5659703	63.44
SmartScreen	3177011	35.61
OrganizationIdentifier	2751518	30.84
SMode	537759	6.03
CityIdentifier	325409	3.65
Wdft_IsGamer	303451	3.40
Wdft_RegionIdentifier	303451	3.40
Census_InternalBatteryNumberOfCharges	268755	3.01
Census_FirmwareManufacturerIdentifier	183257	2.05

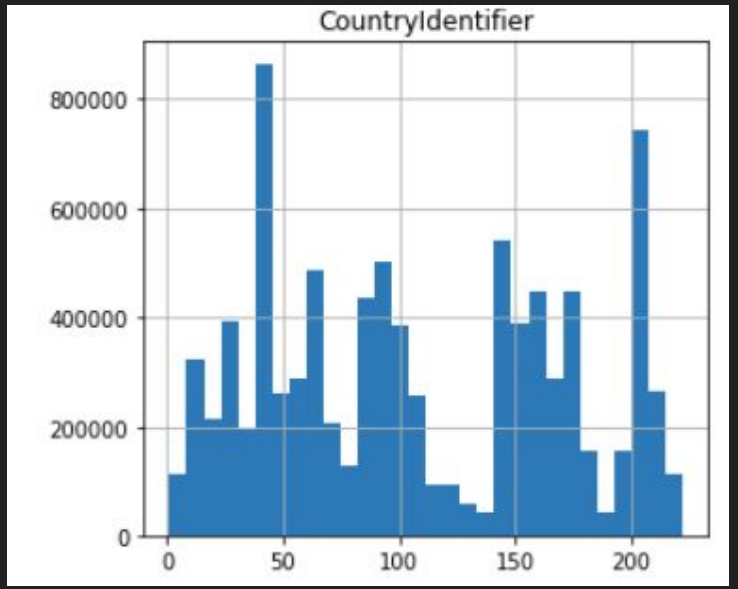
- Dropped features with more than 40% missing values
- No duplicates founds

Data Wrangling - Numerical data

- Errors in records converted to null values
- Skewed distributions and obvious outliers

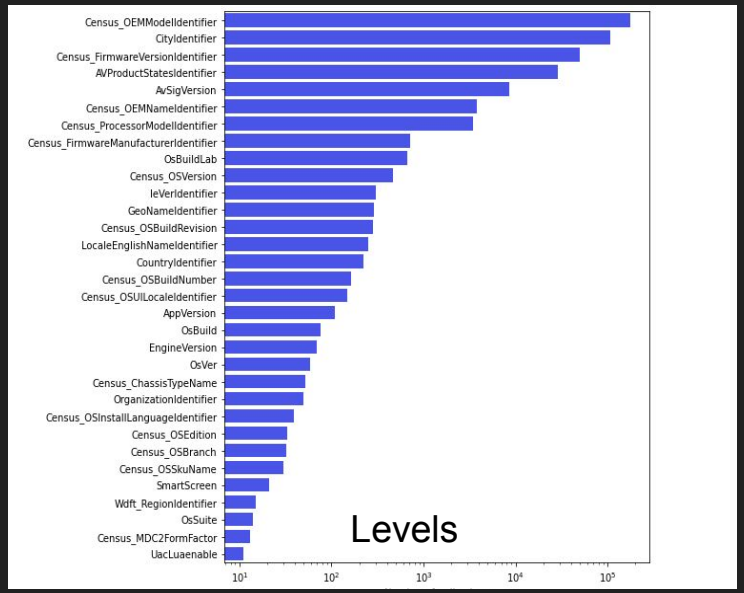


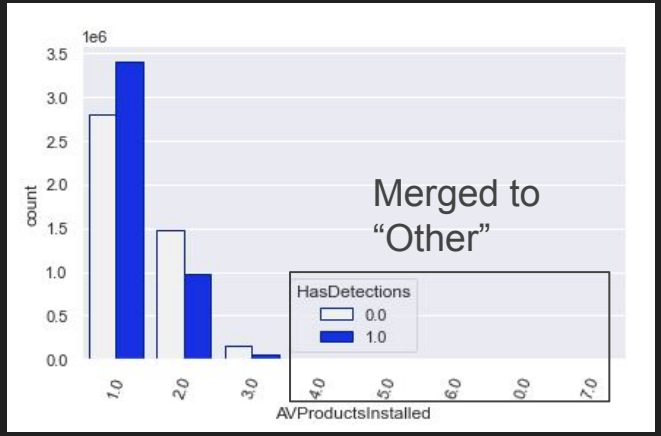
- ID features numerically encoded converted to category dtype



Data Wrangling - Categorical Data

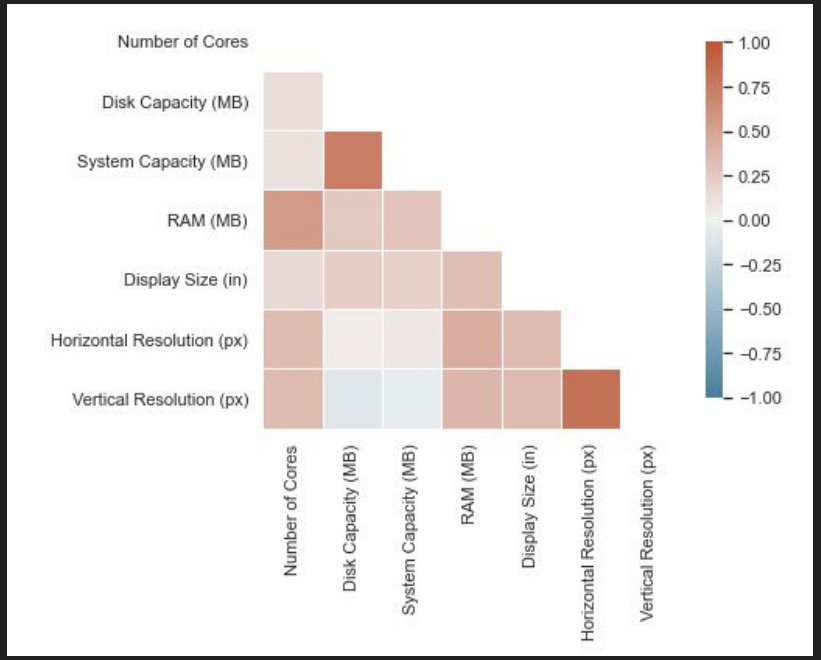
- Spelling typos and inconsistencies: categories labeled as "OFF", "Off", "off"
- Reduced cardinality of categorical features (ID columns excluded)
- Redundant OS and Census features dropped: OsBuildLab, Census_OSBuildNumber

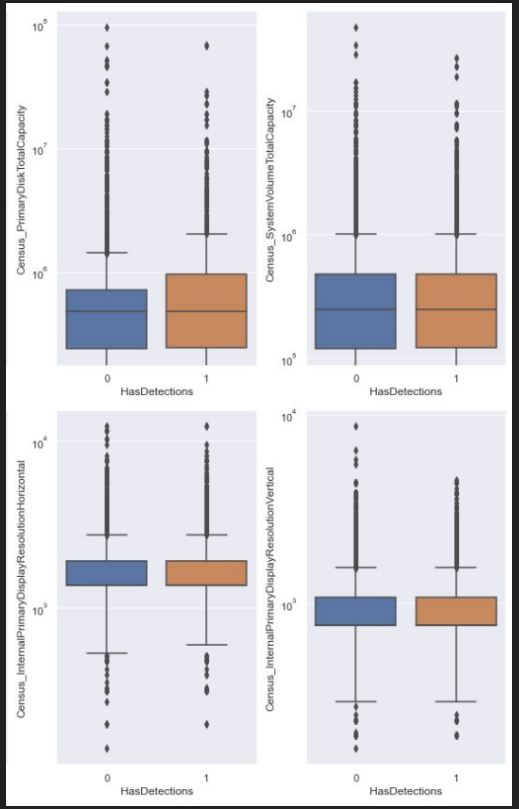




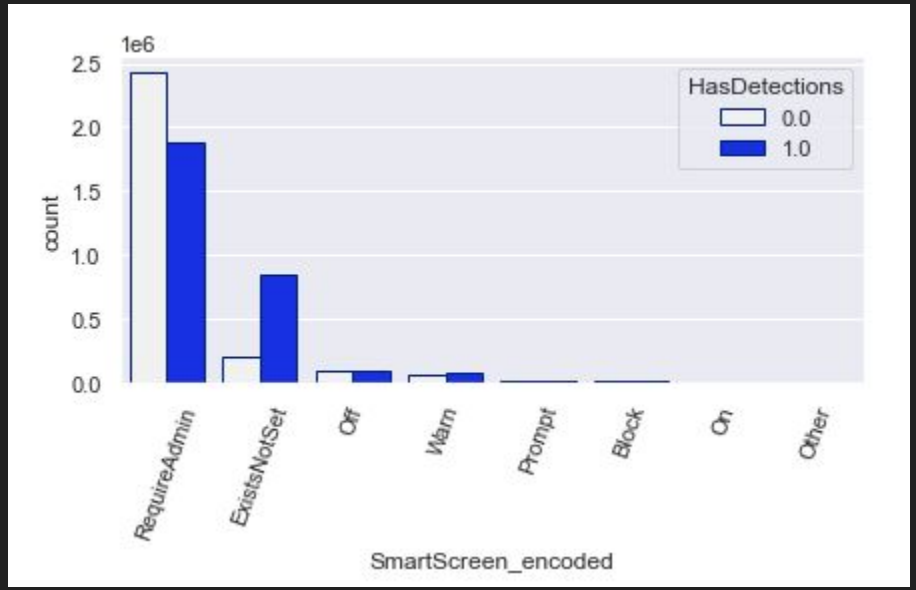
Exploratory Data Analysis - Numerical Data

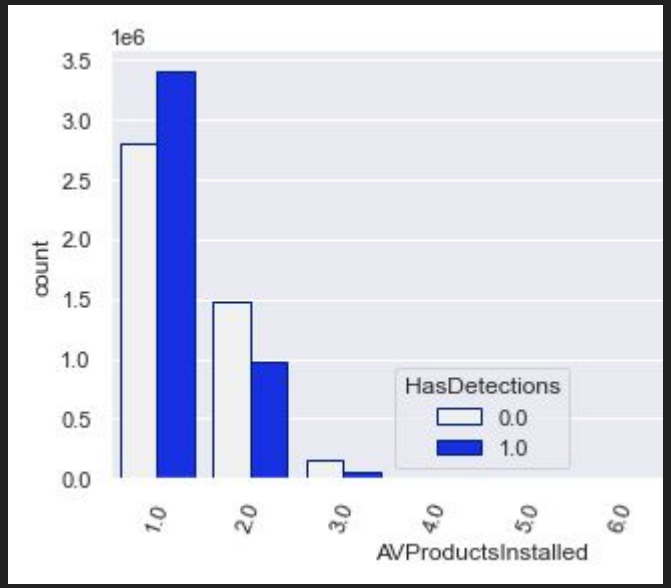
Spearman's Rank Correlation



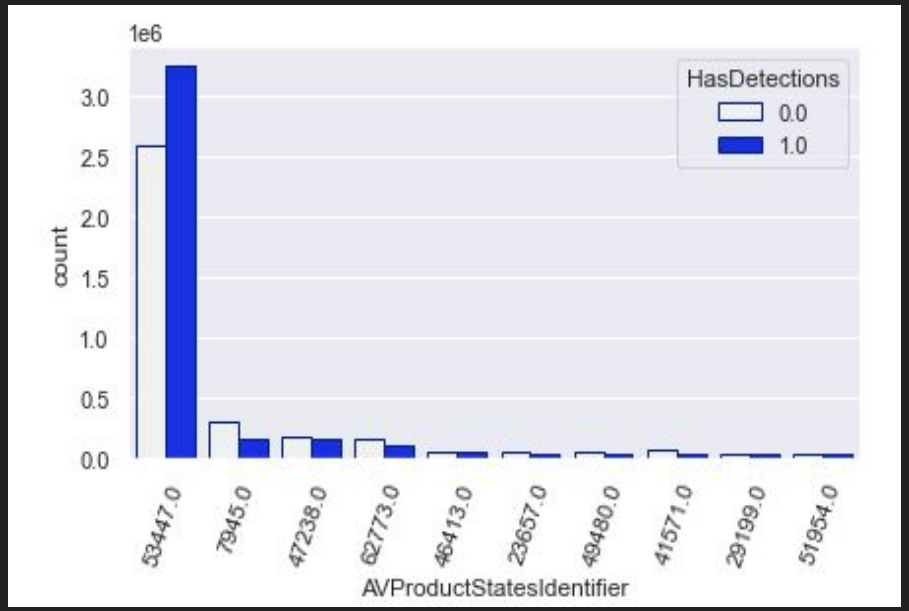


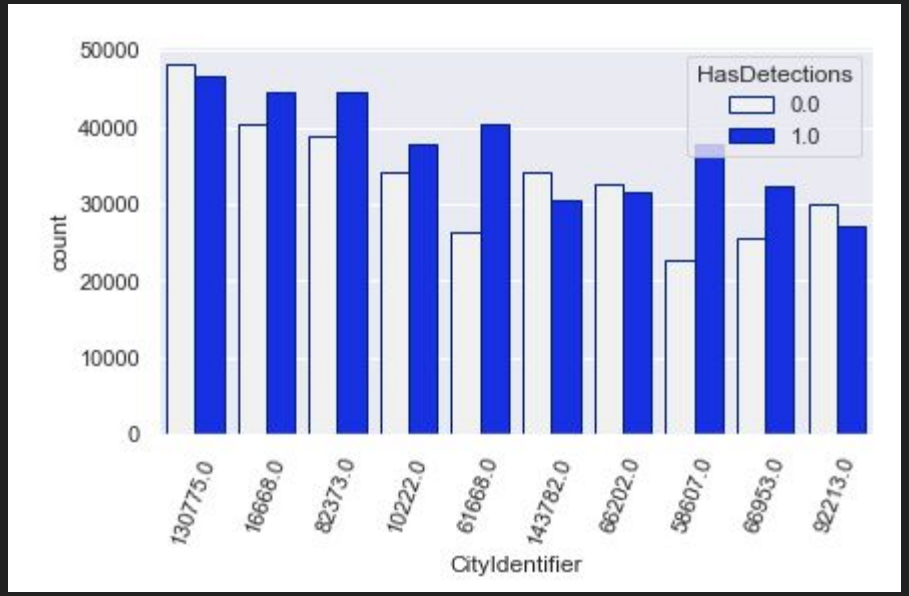
Exploratory Data Analysis - Categorical Data



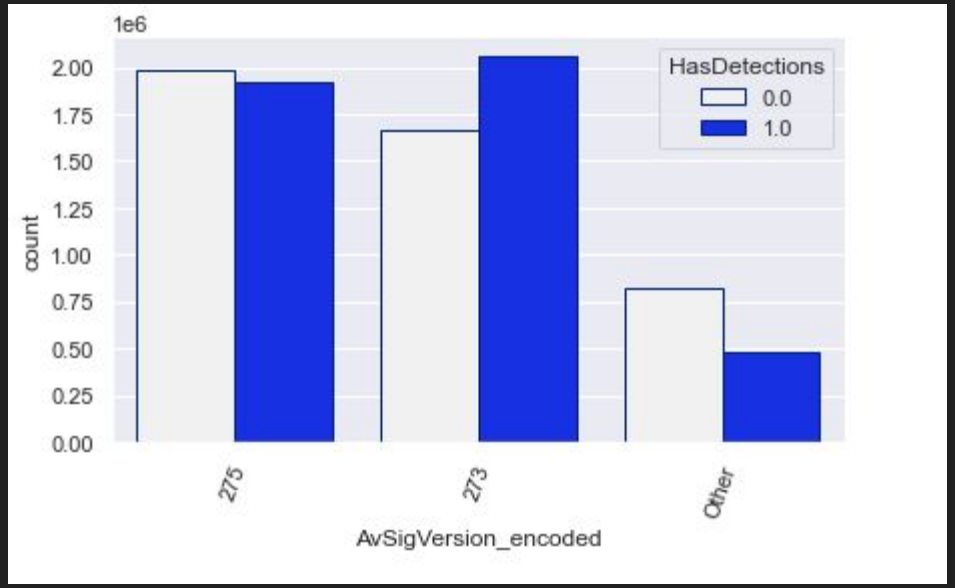


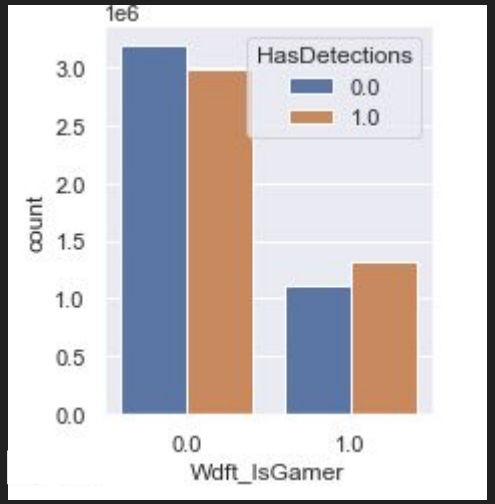
Exploratory Data Analysis - Categorical Data





Exploratory Data Analysis - Categorical Data

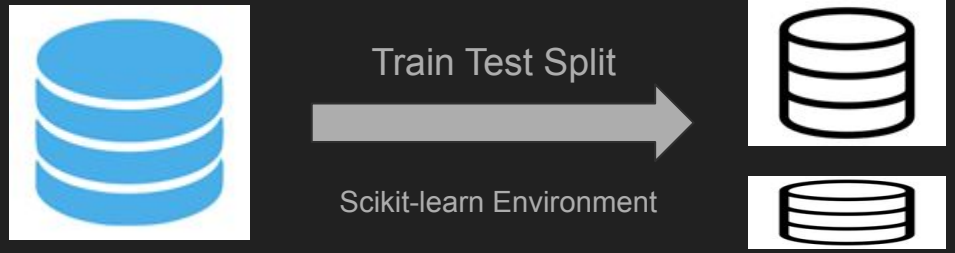




Top 20 features by chi2 statistics

	Chi2-statistics	p-value
SmartScreen_encoded_ExistsNotSet	2.214759e+05	0.000000e+00
SmartScreen_encoded_RequireAdmin	5.698300e+04	0.000000e+00
AvSigVersion_encoded_Other	4.569324e+04	0.000000e+00
AVProductsInstalled_2.0	4.478073e+04	0.000000e+00
AVProductsInstalled_1.0	2.998999e+04	0.000000e+00
Processor_x86	2.574212e+04	0.000000e+00
Census_OSArchitecture_x86	2.557563e+04	0.000000e+00
AvSigVersion_encoded_273	2.183277e+04	0.000000e+00
EngineVersion_encoded_15100	2.160499e+04	0.000000e+00
EngineVersion_encoded_14901	1.711393e+04	0.000000e+00
Census_PowerPlatformRoleName_Slate	1.686514e+04	0.000000e+00
AppVersion_encoded_14	1.560553e+04	0.000000e+00
AVProductsInstalled_3.0	1.482473e+04	0.000000e+00
IsProtected_1.0	1.297864e+04	0.000000e+00
AppVersion_encoded_16	1.270238e+04	0.000000e+00
EngineVersion_encoded_15000	1.224884e+04	0.000000e+00
AppVersion_encoded_18	1.111389e+04	0.000000e+00
EngineVersion_encoded_14800	1.027182e+04	0.000000e+00
AVProductsEnabled_2.0	9.696989e+03	0.000000e+00
Census_MDC2FormFactor_encoded_Detachable	8.591842e+03	0.000000e+00

Feature Engineering



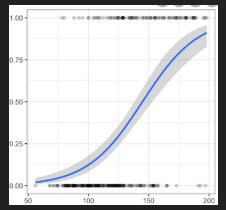
Preprocessing Pipeline

- Missing values imputation
- Categorical variables One Hot Encoding
- ID variables Target Encoding
- 511 columns obtained

```
# Preprocessing for numerical data
numerical_transformer = SimpleImputer(strategy='median')

# Preprocessing for binary data
binary_transformer = SimpleImputer(strategy='most_frequent')

# Preprocessing for categorical data encoded as numerical ID's
id_transformer = Pipeline(steps=[
    ('encoding', TargetEncoder())
])

# Preprocessing for categorical data
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])

# Bundle transformers
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, num_col),
        ('bin', binary_transformer, binary_col),
        ('id', id_transformer, id_col),
        ('cat', categorical_transformer, cat_col)
    ])
```

Modeling

- Logistic Regression with LASSO Regularization
- Ensemble Models, Bagging and Boosting Models
  - Random Forests
  - XGBoost, LightGBM
  - Not as interpretable but offer feature importances

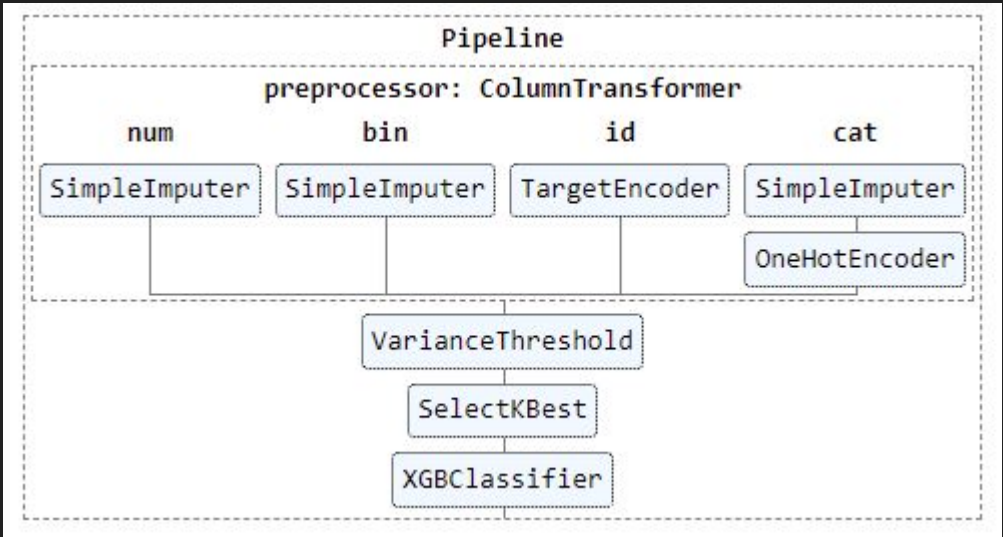


XGBoost

LightGBM

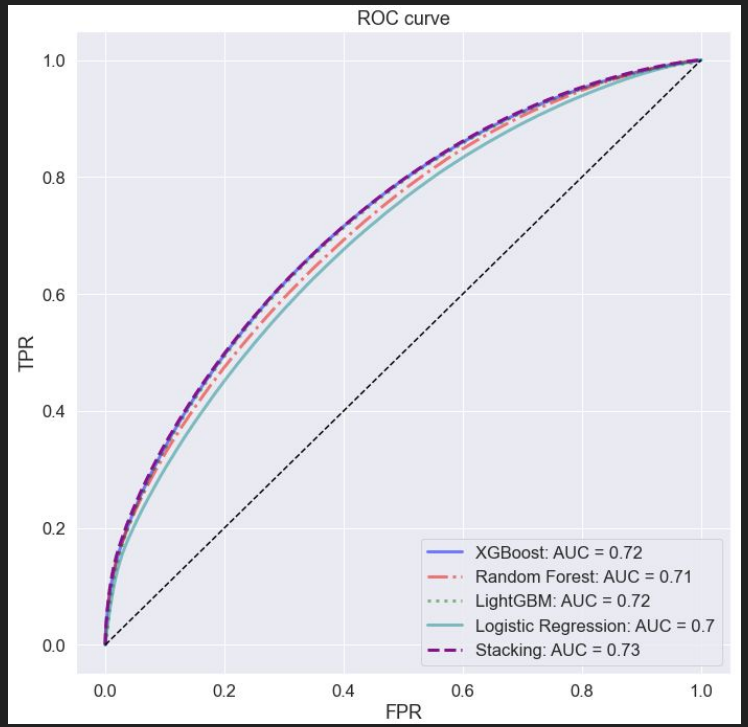
Hyperparameter Tuning and Pipeline

- Tuning and Overfitting addressed with Randomized Search:
  - 3-fold-cross-validation and 15 iterations due to limitations in computational power and time constraints
  - 5-fold-cv and 60 iterations advisable if computational power allows it



Performance Evaluation on Test Set

Explored Stacked model to boost performance

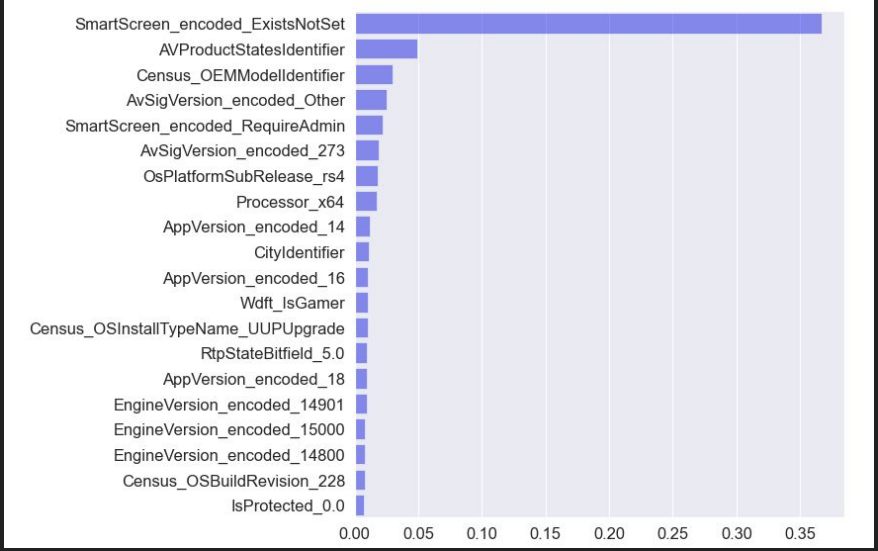


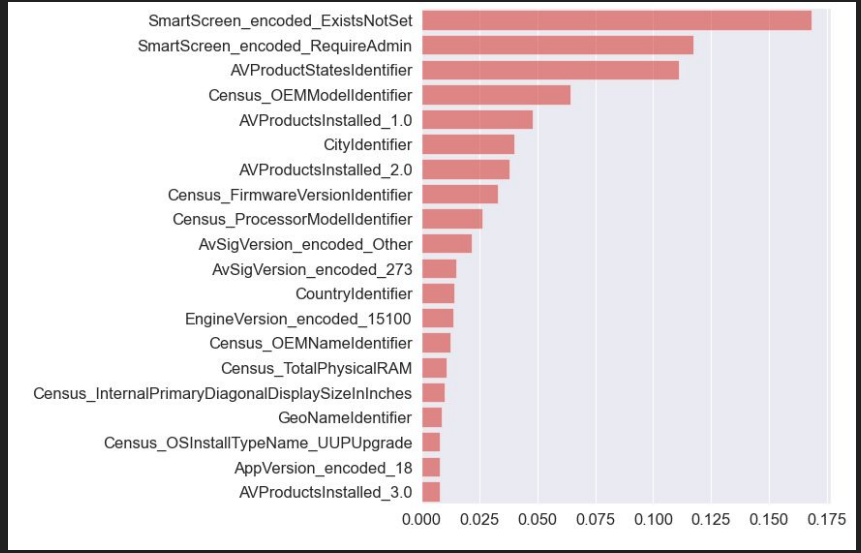
	Coefficient
Model	
XGBoost	2.169014
Random Forest	3.303607
LightGBM	1.313139
Logistic Regression	-1.748189

	precision	recall	f1-score	support
0	0.66	0.66	0.66	892394
1	0.66	0.66	0.66	891902

Balanced data, balanced performance
Threshold may be adapted to increase recall

Feature Importance confirmed EDA findings





	Chi2-statistics	p-value
SmartScreen_encoded_ExistsNotSet	2.214759e+05	0.000000e+00
SmartScreen_encoded_RequireAdmin	5.698300e+04	0.000000e+00
AvSigVersion_encoded_Other	4.569324e+04	0.000000e+00
AVProductsInstalled_2.0	4.478073e+04	0.000000e+00
AVProductsInstalled_1.0	2.998999e+04	0.000000e+00
Processor_x86	2.574212e+04	0.000000e+00
Census_OSArchitecture_x86	2.557563e+04	0.000000e+00

Future Work

- Approach the problem with a different “temporal” strategy
  - Malware infection determined by how up-to-date the machine protection is
  - Temporal sampling is important
- Stratified models an option to explore, e.g. for different OS/AV/Defender versions
- Adopt parallel computing solutions to investigate a wider range of hyperparameters

Thank you!