Predicting Malware Infection of Windows Machines

Pavel Zimin, PhD

Mentor: Ramkumar Hariharan, PhD

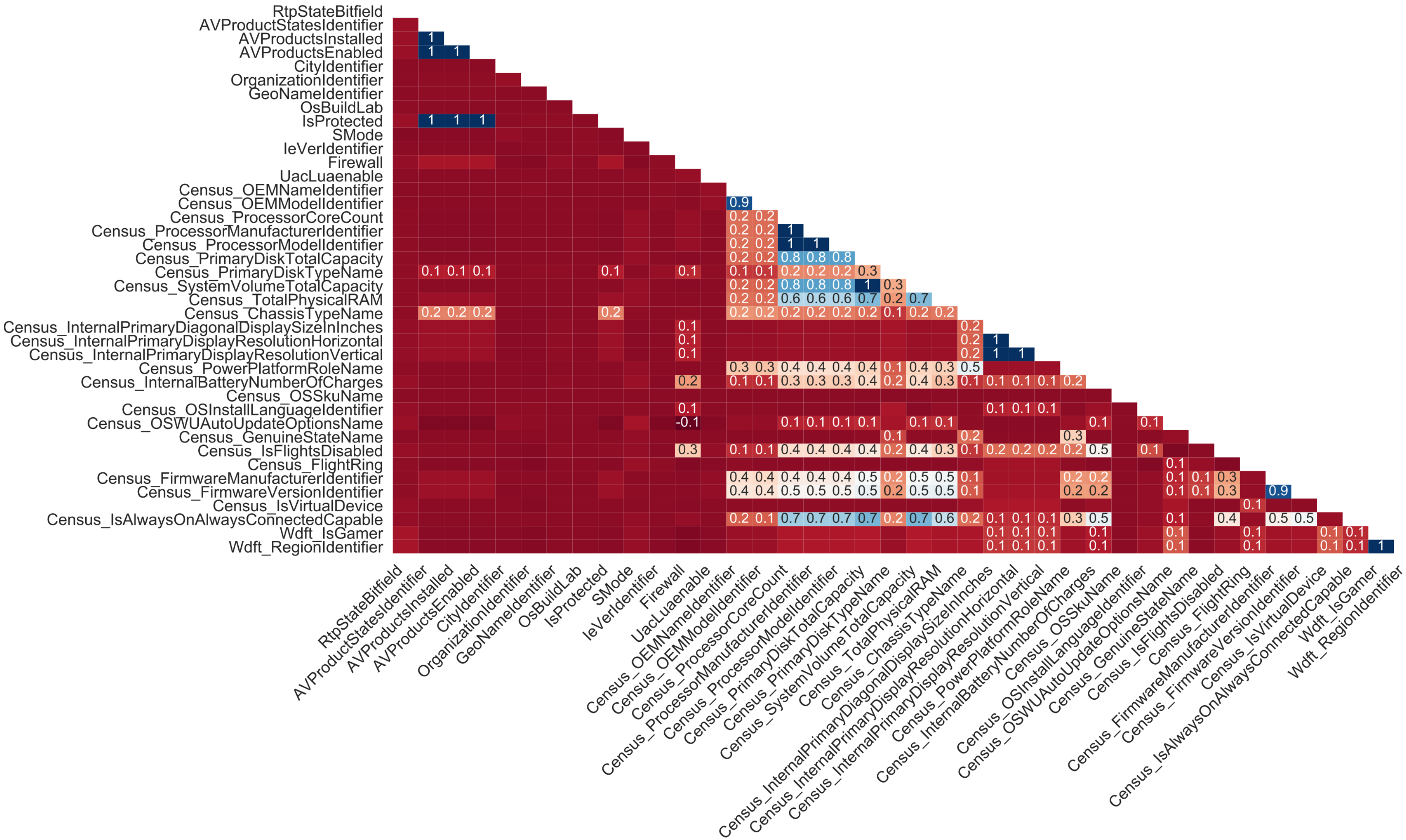
Problem Statement

- Computer infection by malware constitutes a serious security problem
- The ability to predict the chances of malware infection before they occur would benefit consumers and businesses

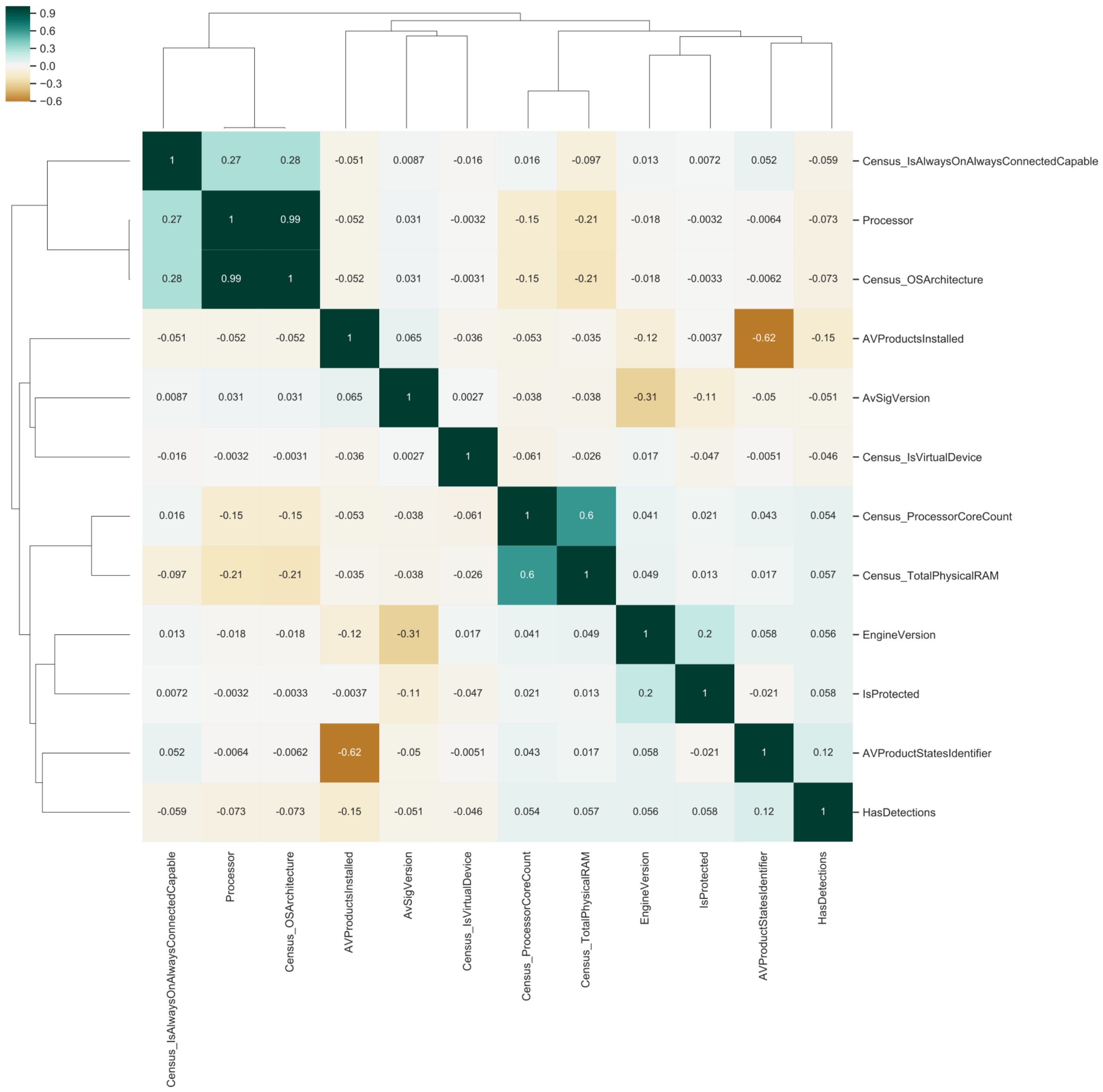
Business Use Scenario:

- This project would benefit software manufacturers who would be able to incorporate the model into their software that would allow for additional security measures aimed at preventing the infection by malware

Correlation heatmap for missing values



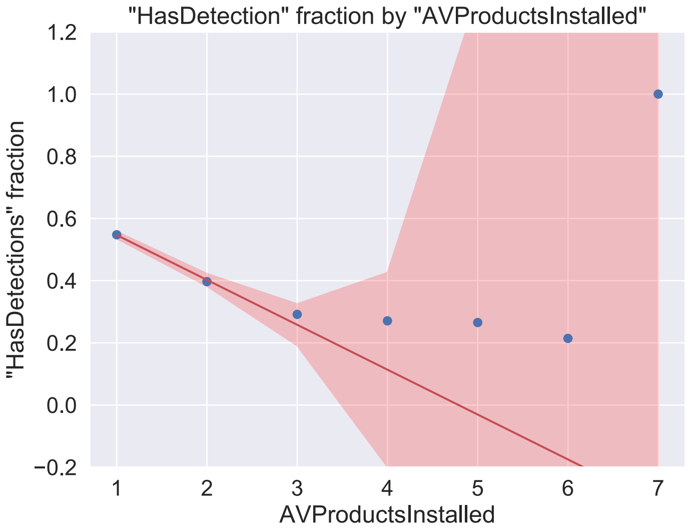
Correlation Cluster Heatmap

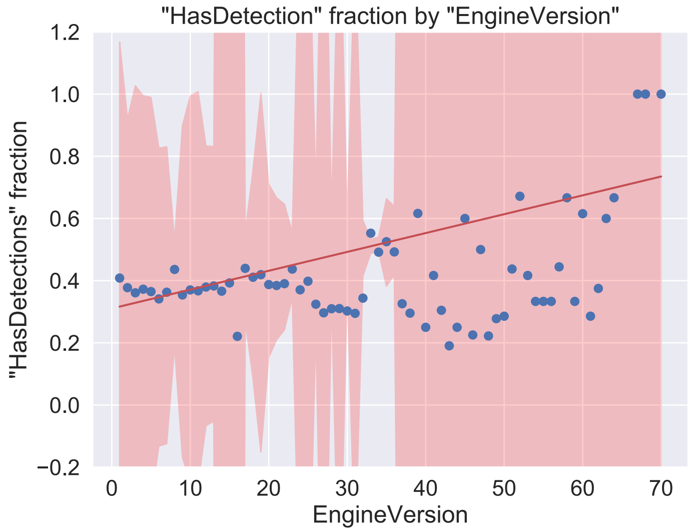


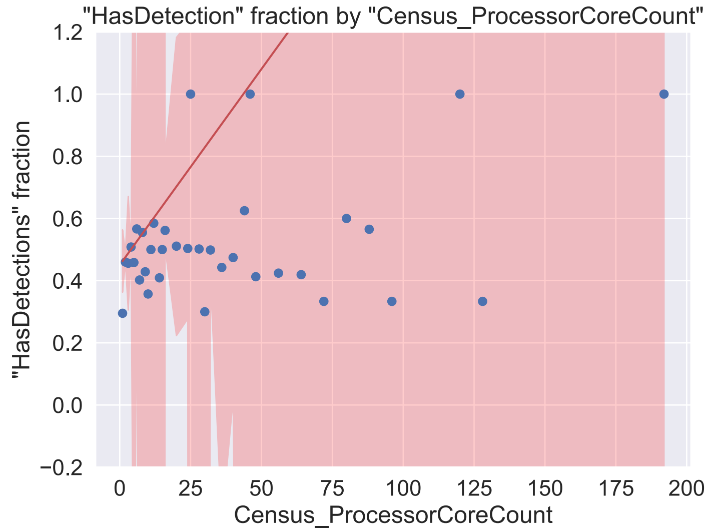
	Census_IsAlwaysOnAlwaysConnectedCapable	Processor	Census_OSArchitecture	AVProductsInstalled	AvSigVersion	Census_IsVirtualDevice	Census_ProcessorCoreCount	Census_TotalPhysicalRAM	EngineVersion	IsProtected	AVProductStatesIdentifier	HasDetections
Census_IsAlwaysOnAlwaysConnectedCapable	1	0.27	0.28	-0.051	0.0087	-0.016	0.016	-0.097	0.013	0.0072	0.052	-0.059
Processor	0.27	1	0.99	-0.052	0.031	-0.0032	-0.15	-0.21	-0.018	-0.0032	-0.0064	-0.073
Census_OSArchitecture	0.28	0.99	1	-0.052	0.031	-0.0031	-0.15	-0.21	-0.018	-0.0033	-0.0062	-0.073
AVProductsInstalled	-0.051	-0.052	-0.052	1	0.065	-0.036	-0.053	-0.035	-0.12	-0.0037	-0.62	-0.15
AvSigVersion	0.0087	0.031	0.031	0.065	1	0.0027	-0.038	-0.038	-0.31	-0.11	-0.05	-0.051
Census_IsVirtualDevice	-0.016	-0.0032	-0.0031	-0.036	0.0027	1	-0.061	-0.026	0.017	-0.047	-0.0051	-0.046
Census_ProcessorCoreCount	0.016	-0.15	-0.15	-0.053	-0.038	-0.061	1	0.6	0.041	0.021	0.043	0.054
Census_TotalPhysicalRAM	-0.097	-0.21	-0.21	-0.035	-0.038	-0.026	0.6	1	0.049	0.013	0.017	0.057
EngineVersion	0.013	-0.018	-0.018	-0.12	-0.31	0.017	0.041	0.049	1	0.2	0.058	0.056
IsProtected	0.0072	-0.0032	-0.0033	-0.0037	-0.11	-0.047	0.021	0.013	0.2	1	-0.021	0.058
AVProductStatesIdentifier	0.052	-0.0064	-0.0062	-0.62	-0.05	-0.0051	0.043	0.017	0.058	-0.021	1	0.12
HasDetections	-0.059	-0.073	-0.073	-0.15	-0.051	-0.046	0.054	0.057	0.056	0.058	0.12	1

- The figure shows only the features with the highest positive or negative Pearson correlation coefficients with the target variable
- Overall low correlation of features with the target variable
- The following features show the highest correlation with target variable:
  - ‘AVProductsInstalled’
  - ‘AVProductStatesIdentifier’
  - ‘Processor’
  - ‘Census_OSArchitecture’

Frequencies of target variable by selected features



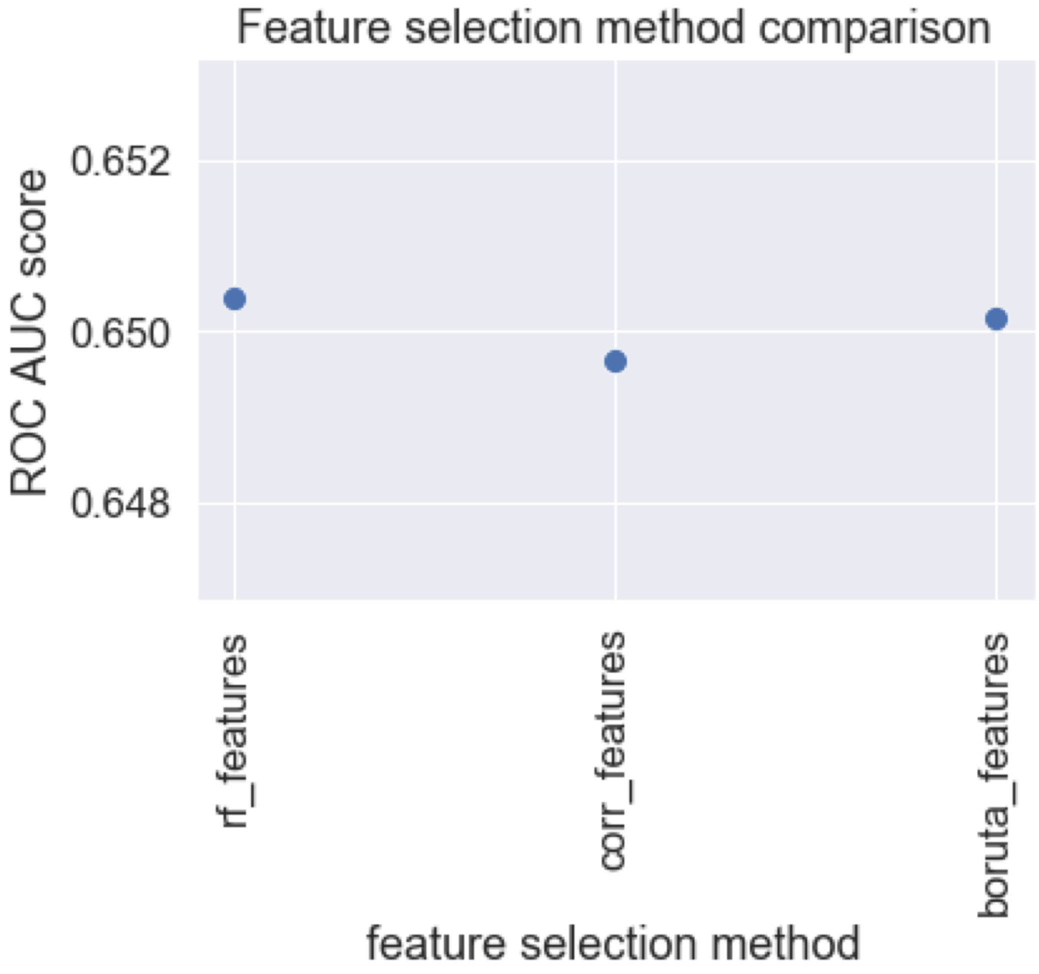




Weighted least squared models were fitted to the data with the weights of the value count for each data point.

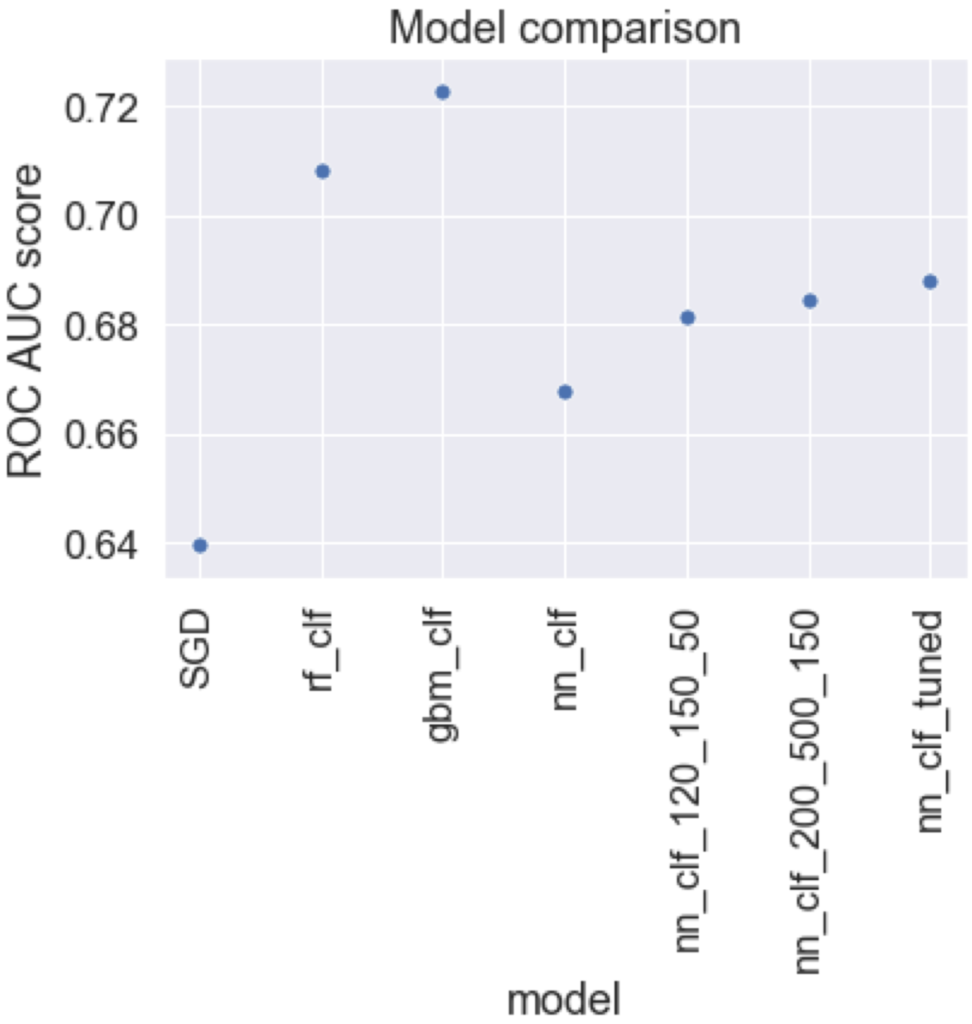
Feature Selection

- 3 methods of feature selection were performed:
  - Random Forest’s feature importances
  - Elimination of highly correlated features
  - Boruta algorithm



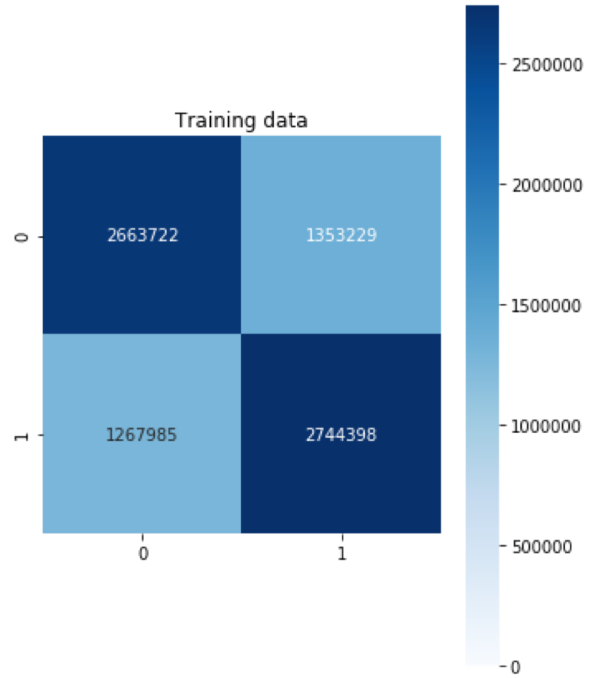
Model Selection

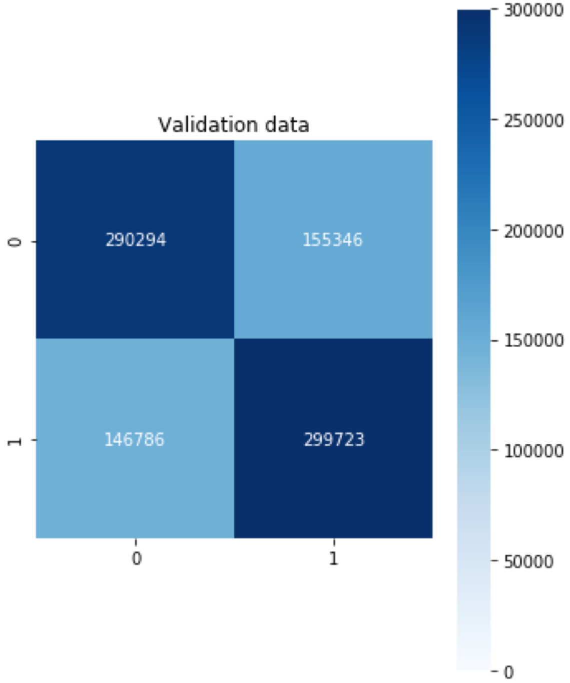
- Logistic Regression (SGD)
- Random Forest (rf_clf)
- Gradient Boosting (gbm_clf)
- Neural networks (nn_)

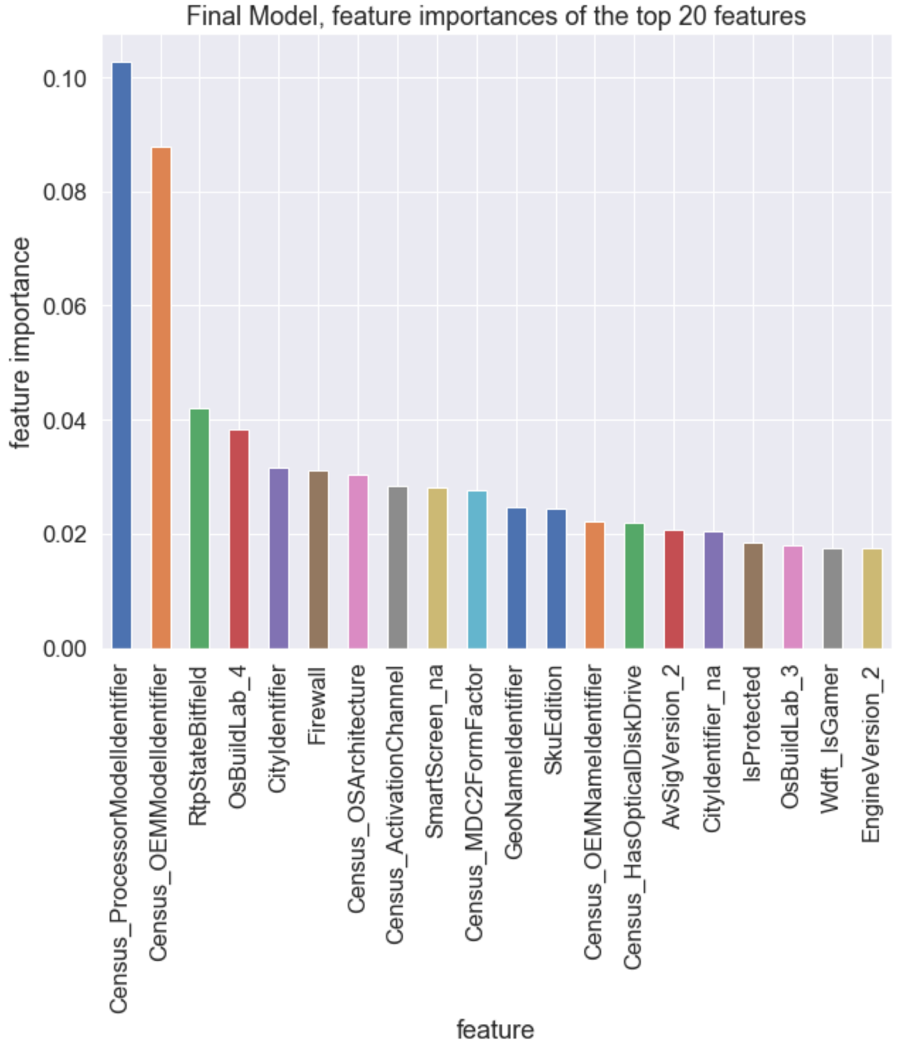


Final Model Evaluation and Feature Importances

- ROC AUC score 0.74 on test data







Conclusions

- Related features tend to have the same observation with missing values.
- Features show low overall correlation with the target variable.
- Weighted least squared models show significant relationship between selected features and the target variable.
- Features were selected based on random forest’s feature importances.
- Best performing model was Gradient Boosting with ROC AUC score 0.74.
- The most important features are: Census_ProcessorModelIdentifier, Census_OEMModelIdentifier, RtpStateBitfield, OsBuildLab_4, CitiIdentifier, Firewall.