Predicting Malware Infection of Windows Machines

Pavel Zimin, PhD

Mentor: Ramkumar Hariharan, PhD

Problem Statement

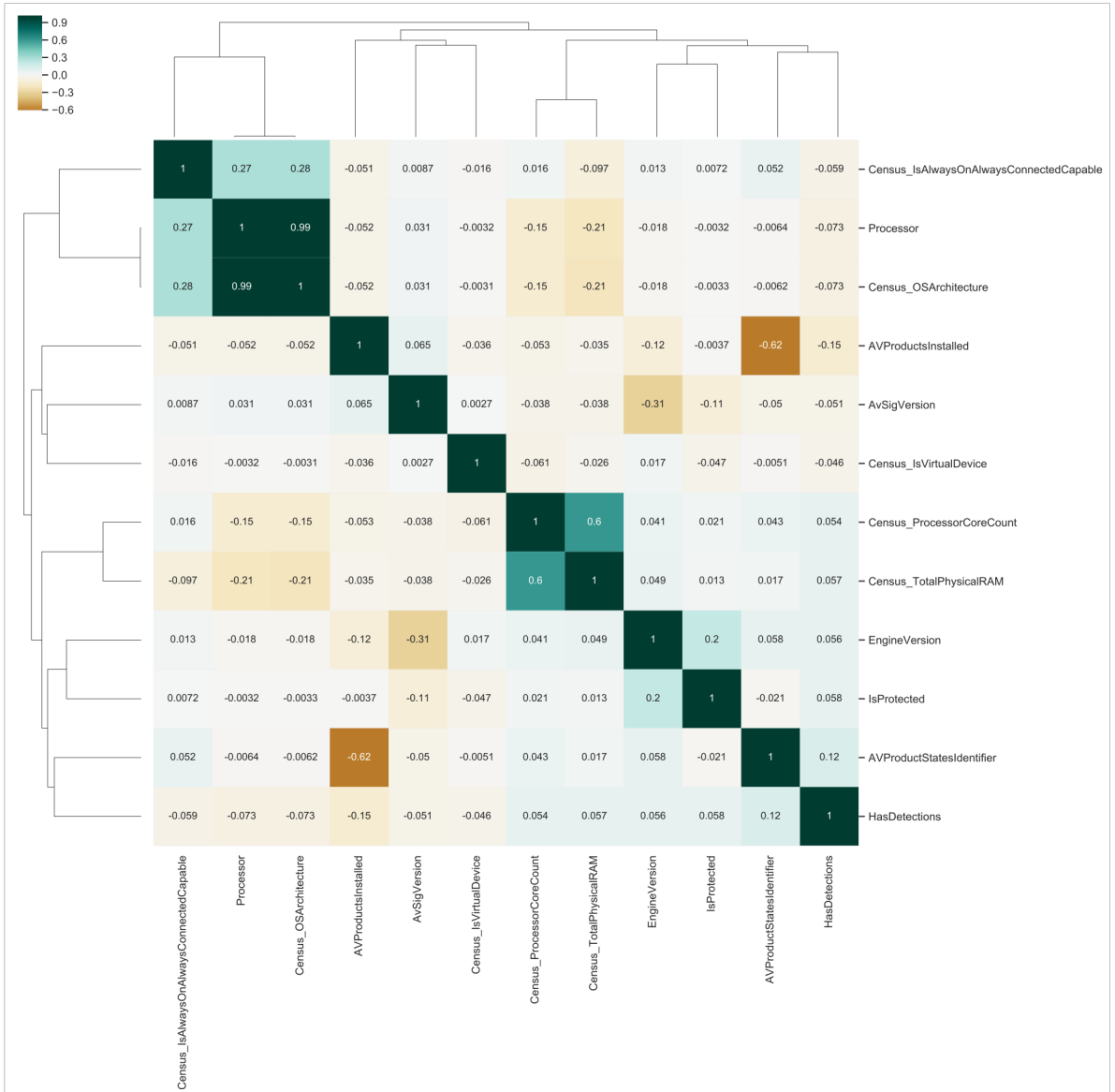
- Computer infection by malware constitutes a serious security problem
- The ability to predict the chances of malware infection before they occur would benefit consumers and businesses

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

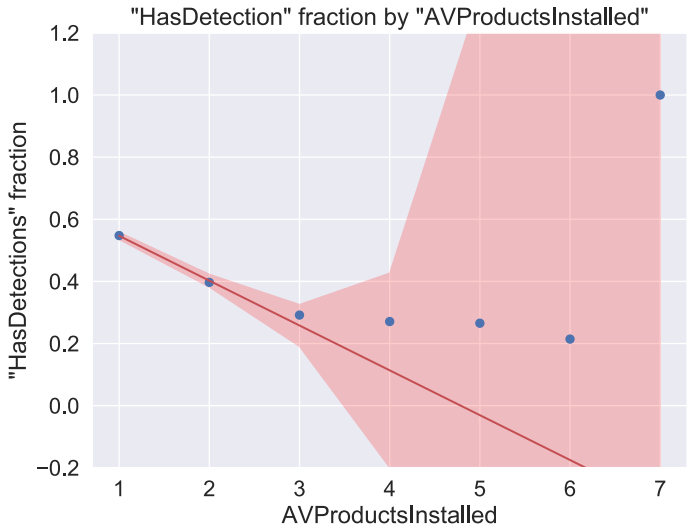
Correlation heatmap for missing values

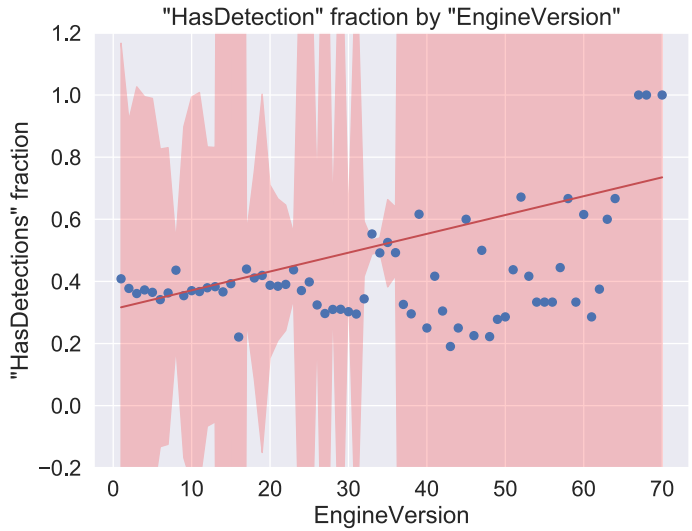
Correlation Clustermap

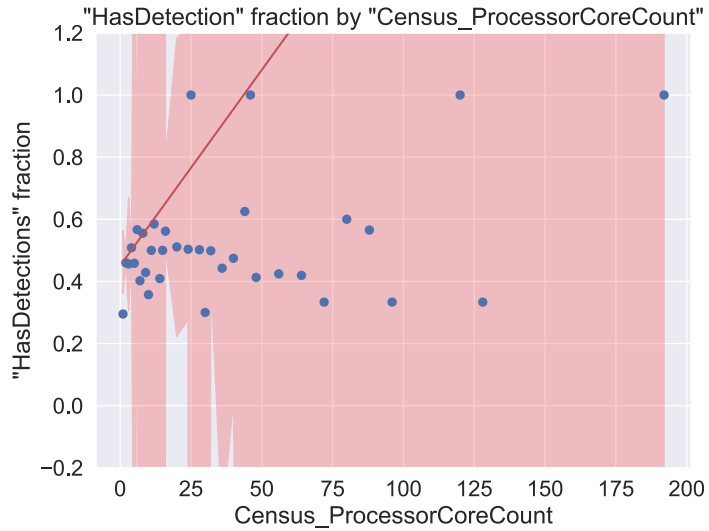


- Overall low correlation of features with the target variable
- The following features show the highest correlation with target variable:
  - ‘AVProductsInstalled’
  - ‘AVProductStatesIdentifier’
  - ‘Processor’
  - ‘Census_OSArchitecture’

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.

Conclusions

- Related features tend to have the same observation with missing values.
- Features show low overall correlation with the target variable.
- The features that show the highest correlation coefficients with the target variable are ‘AVProductsInstalled’, ‘AVProductStatesIdentifier’, ‘Processor’, ‘Census_OSArchitecture’.
- Weighted least squared models show significant relationship between selected features and the target variable.