Microsoft Malware Prediction

Authors: Anchal Atlani, Anushka Kamalapurkar, Anup Bharade, Apurva Nanajkar, Mandar Phapale, Rithesh Shetty, Omkar Marne, Nikhil Jadhav, Sagar Kulkarni, Shravanti Bugade, Soham Dhodapkar

University of North Carolina at Charlotte

\*Contact: [aatlani@uncc.edu](mailto:aatlani@uncc.edu), [akamala1@uncc.edu](mailto:akamala1@uncc.edu), [abharade@uncc.edu](mailto:abharade@uncc.edu), [ananajka@uncc.edu](mailto:ananajka@uncc.edu), [mphapale@uncc.edu](mailto:mphapale@uncc.edu), [rshetty2@uncc.edu](mailto:rshetty2@uncc.edu), [omarne@uncc.edu](mailto:omarne@uncc.edu), [njadhav@uncc.edu](mailto:njadhav@uncc.edu), [skulka23@uncc.edu](mailto:skulka23@uncc.edu), [sbugade@uncc.edu](mailto:sbugade@uncc.edu), [sdhodapk@uncc.edu](mailto:sdhodapk@uncc.edu)

Abstract - ***Malware is a standout amongst the most genuine security dangers and spreads autonomously through vulnerabilities or lack of regard of clients.[1] Accurately detecting a malware helps us to protect a computer from getting affected. Detection of a malware is generally based on methods and techniques which help us efficiently determine the class of malware. As we are familiar with numerous theoretical challenges, detection of malware also has practical challenges viz. the speed at which the malwares are developed and distributed and well developed techniques for escaping polymorphism and metamorphism detection.***

***In this Paper, we have taken a dataset provided by Microsoft which has unprecedented malware data in tabular format. We have shown predictions using different Analysis techniques like Classification, Clustering, etc. We have used feature selection and dimensionality reduction techniques to eliminate the features which are not required for further analysis. We intend to check the prediction accuracy and error rate of our prediction to analyze which model is most suitable for this data.***

***Keywords— Malware, XLMiner, Data Visualization, Prediction, Regression, Clustering, Association***

1. Introduction

The Internet has lowered the development time and extended the distribution speed of latest malware by various orders of measurement. This improvement has encouraged the active release of adjusted and refined malware modifications in short time. Within very less time many computer systems gets infected by malware despite of good anti-virus protection. One of the stated aims of research in malware detection is to reduce or eradicate this window of vulnerability. A complete set of statistical and machine learning based methods is provided by XLMiner. A problem or a data set can be examined by several means. Trying various approaches by comparing their results and then choosing best model usually suits the problem well. XLMiner can operate with large data sets which many of the times not possible in Excel [2].

Our data set comprises of 50,000 rows and 82 columns having different ranges and frequencies. The second point explains the business use case and the third point is about the hypothesis. The fourth point elaborates the information about the dataset and the pre-processing of data. In the fifth point, classification and prediction models are applied to the data based on the hypotheses. A set of three models are applied on two hypothesis and then the reports are validated to select the best model based on various measures such as accuracy, lift charts, error percentage, and others. The models used to address the hypotheses are K-means Clustering, Classification Tree, Linear Regression, Logistic Regression, KNN Regression. The models developed using these techniques would help us analysis which model is the most suitable for this type of data.

1. BUSINESS USE CASES

Malware detection in a system is critically important pre inference of any security system. Data mining and modelling techniques can be very useful to predict which systems are more vulnerable to a Malware based on various features of the systems. On finding out the highest probable features to get affected, a company can further drill down to check what kind of machines have more chance of getting affected, and then shut those machines down or upgrade their security specifications.

Certain security flaws or software bugs could largely influence vulnerability into a product. Modelling and predicting the target machines can help a company push software updates or releases to such highly vulnerable machines.

1. HYPOTHESIS QUESTIONS

* Out of all the variables, which ones are the most useful for Malware detection?
* How are gaming devices being affected by malwares?
* Which type of device is most prone to Malware attacks?

1. Data set

The Microsoft data set has been taken from Kaggle. Each observation in this dataset belongs to a machine which is uniquely identified by a MachineIdentifier. HasDetections is the output variable truth which indicates that Malware was detected on the machine. In this the goal is to use the training data build a model which can be used to predict the Malware presence in any machine given its characteristics.

The dataset provided by Microsoft has sampled according to their business constraint considering the user privacy and the time at which the machine was being used. Malware detection is the time series problem; however, this time series problem becomes more problematic by addition of new machines or by machines that comes online and offline etc. This dataset is split by time which also means that there might flawed agreement between scores, cross-validation etc. Moreover, this dataset has not been taken randomly nor is it the representative of the Microsoft customers it has been filtered in such a way that it constitutes a larger proportion of the malware systems.

The description of the various data fields is:

* MachineIdentifier - unique machine identifier assigned to each machine.
* ProductName - Defender state information e.g. win8defender
* EngineVersion - Defender state information e.g. 1.1.15100.1
* AppVersion - Defender state information e.g. 4.18.1807.18075
* AvSigVersion - Defender state information e.g. 1.273.1527.0
* IsBeta - Defender state information e.g. true
* CountryIdentifier – country identifier where the machine is located
* CityIdentifier – City identifier where the machine is located.
* OrganizationIdentifier – This is the identifier which gives information about the organization a machine belongs to.
* GeoNameIdentifier – This variable gives the information about the geographic region of machine.
* LocaleEnglishNameIdentifier – Locale Id name of the user in English.
* Platform - Computes stage name (of OS related properties and processor property)
* Processor - This variable gives the information about the processor architecture of the machine operating system.
* OsVer – This variable signifies the Operating System version
* OsBuild - This variable signifies the Operating System Build
* OsSuite - Product suite mask for the current operating system.
* OsPlatformSubRelease – This variable gives the information of sub release version of the operating system.(Windows Vista, Windows 7, Windows 8, TH1, TH2)
* OsBuildLab - Build lab that generated the current OS. Example: 17134.1.amd64fre.rs4\_release.180410-1804
* SkuEdition - The objective of this element is to utilize the Product Type characterized in the MSDN to guide to a 'SKU-Edition' name that is valuable in populace detailing. The legitimate Product Type are characterized in %sdxroot% information windowseditions.xml. This API has been utilized since Vista and Server 2008, so there are numerous Product Types that don't make a difference to Windows 10. The 'SKU Edition' is a string esteem that is in one of three classes of results. The plan must hand each class.
* IsProtected – This variable gives information about the antivirus. It is a categorical variable can take following values. Genuine esteem tells that in any event one dynamic and forward-thinking antivirus item running on this machine. FALSE esteem tells if there is no dynamic AV item on this machine, or if the AV is dynamic, yet isn't getting the most recent updates, and invalid esteem tells if there are no Anti-Virus Products. This tells whether a machine is protected or not.
* AutoSampleOptIn - This is the SubmitSamplesConsent esteem gone in from the administration, accessible on CAMP 9+
* SMode – This variable tells about the ’S Mode’ of machine. It is a categorical variable. If the value of this variable is true, then the device is known to be in ’S Mode’. For instance, Windows 10 S mode, where just Microsoft Store applications can be introduced.
* SmartScreen - This is the SmartScreen enabled or not string. This value is obtained from the HKLM Software policy of the machine. On the off chance that the esteem exists however is clear, the esteem "ExistsNotSet" is sent in telemetry.
* Firewall – This variable gives information about firewall enabled or not. The value is set to 1 for Windows 8.1 and above if windows firewall enabled.
* Census\_MDC2FormFactor - A gathering dependent on a blend of Device Census level equipment attributes. The rationale used to characterize Form Factor is established in business and industry models and lines up with how individuals consider their gadget. (Examples: Smartphone, Small Tablet, All in One, Convertible...)
* Census\_DeviceFamily – This variable gives the type of device classs. Indicates the type of device that an edition of the OS is intended for. Example values: Windows.Desktop, Windows.Mobile, and iOS.Phone.

*A. Data Pre-Processing*

The dataset contains total 82 number of columns, out of which some columns contain redundant data and some columns contain null values, also few columns are highly correlated. We have performed cleaning and pre-processing of data which includes activities like cleaning the data by either removing null values columns or by replacing it with a logical value.

We have removed PuaMode, Census\_ProcessorClass, DefaultBrowsersIdentifier, Census\_IsFlightingInternal and Census\_InternalBatteryType, Census\_IsWIMBootEnabled, Census\_ThresholdOptIn, Smartscreen organization\_identifier have over 70

Below columns are skewed towards one value and hence we are removing them

* ’Isbeta’,’Census\_IsFlightsDisabled’,’AutoSampleOptIn’, ’Cen-sus\_IsAlwaysOnAlwaysConnectedCapable’, ’IsSxsPassiveMode’,’Census\_HasOpticalDiskDrive’ is skewed towards 0.
* ’SMode’,’Census\_IsPortableOperatingSystem’, ’Census\_IsVir tualDevice’,’Census\_IsPenCapable’ are also skewed on 0. ’Census\_DeviceFamily’ are skwed on values like windows and desktop.
* ’UacLuaenable’,’Firewall’,’AVProductsEnabled’,’IsProtected’, ’HasTpm’ is skewed over value 1.
* ’RtpStateBitfield’ is skewed over value 7. ’Platform’,’Census\_OSSkuName’,’Census\_OSInstallLanguage Identifier’ and ’Processor’ are highly correlated hence we are removing them.
* ’Census\_FlightRing’ is skewed on value Retail.
* ’OsVer’ is skewed in value 10.0.0.0.
* ’ProductName’ is skewed over ’windows defender’.

1. Procedures

Predictive Analytics is concerned with forecasting and statistical modelling to determine the future conceivable outcomes. The hypothesis mentioned above are analyses using the following demonstrating techniques:

1. Clustering

Clustering is the task of dividing the population or data points into several groups such that data points in the same groups are more like other data points in the same group than those in other groups.[3] The end goal is to collate groups with similar characteristics and assign them into clusters. There are two clustering techniques hierarchical and k-means. In hierarchical clustering, clusters are joint to make a significant cluster in an additive manner based on the distance between the clusters and distance between the observations in clusters.

A graph like structure is displayed in this which is called a dendrogram, and this graphical view is used in the analysis. In hierarchical clustering two approaches can be used first start with n clusters and merge them until a single cluster obtained, the second method is opposite of the above start with a single cluster and divide it recursively. The K-means algorithm is a non-hierarchical type of clustering. It is an iterative method that starts with k cluster centers randomly chosen. Every observation is then assigned to the cluster having the closest centroid. After this centroid are again calculated as the addition of new observation or deletion of any observation has changed the centers of the clusters. The reassigning of observations to clusters is repeated until no difference occurs in the next iteration.

*B. Classification*

Classification models are used to predict the categorical target variable for example if we have to predict color of car which is bought by many people of any country; and prediction models are used to predict continuous numerical target. For example, we can build a classification model to categorize animals as either mammal or reptile based on the characteristics they possess, or a prediction model to predict the quarter profit of any multinational company given their revenue and investment details. Classification results are either 0 or 1, or any other category of discrete type. When the target is numerical continuous variables then a regression algorithm is used to build a predictive model, not a classification algorithm.

The classification can be of binary type or multiclass. In binary classification, the output attribute has only two possible values: for example, culprit or innocent. Multiclass targets have more than two values: for example, red, yellow, blue color of cars. When supervised learning is used and model building is done using training data then a classification algorithm finds relationships between the values of the input variables and the values of the output variable. Different classification algorithms use different techniques for finding relationships. The relationships found during this are used as model and this can then be applied to a data set for which the output variable is not known.

*C. Association-Rule*

Association rules are on the off chance that explanations that help to demonstrate the likelihood of connections between information things inside expansive informational collections in different kinds of databases. An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in com- bination with the antecedent. Association rules use the criteria support and confidence to identify the most important relation- ships. Support is an indication of how frequently the items appear in the data. Confidence indicates the number of times the if-then statements are found true. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in customer analytics, market basket analysis, product clustering, catalog design and store layout.

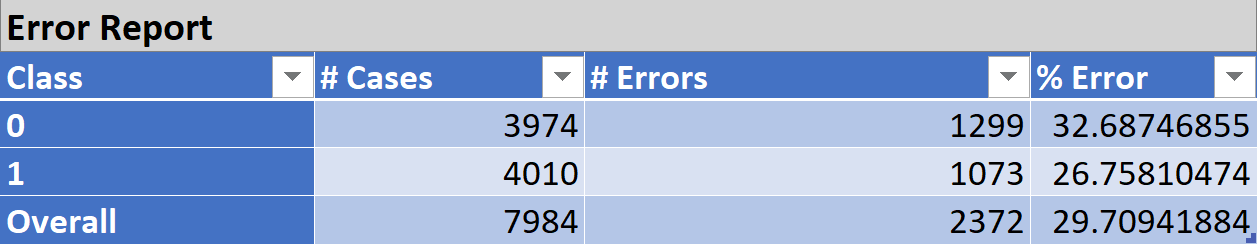
VII. Model Evaluation

* 1. ***Hypothesis 1***

Considering all the variables, or pruned variables from the overall dataset, we will build and compare various prediction model and try to find out which prediction model works best for the given dataset.

***1) Model 1: K Nearest Neighbor Classification:***

K-Nearest Neighbor is one of the simplest Classification methods which uses lazy learning which separates the data points into sev- eral classes to predict the classification of new sample point. The new data point is assigned to a class most common among its k nearest neighbors. It can also be used for regression which gives the output as the value of the object.



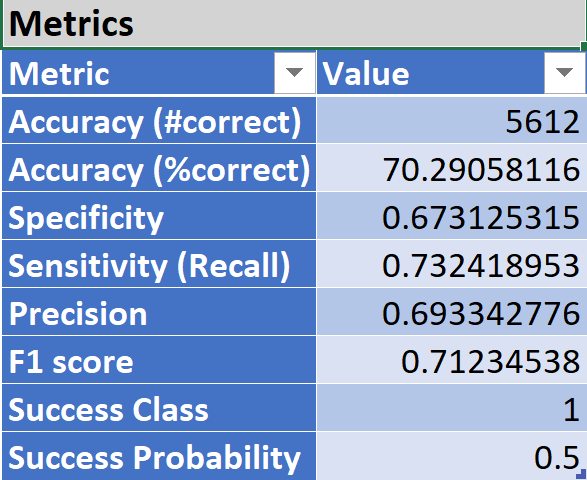
1. KNN Classiftcation error Metrics

Fig. 2. KNN Classification data metrics

***a) Advantages of K Nearest Neighbor***

* KNN does not rely on any underlying assumption about the data.
* KNN is relatively similar than its competitor algorithms.
* KNN has got a good accuracy and can be used for classification as well as regression.

***b) Disadvantages of K Nearest Neighbor***

* The biggest disadvantage of KNN is that it stores all or almost all the training data.
* Because of the above-mentioned reason, it becomes computationally expensive.
* KNN gives high accuracy, but not as compared to the other supervised learning methods.

***c) Best Model***

The Malware Prediction data set has many categorical variables and hence we have chosen the techniques which go with categorical data, KNN, classification, and logistic regression. If we compare the three models applied for this data, then by looking into the metrics the KNN modeling technique has a high accuracy of 70.29 among all three and the error percentage is also lowest. So, for this data set, KNN gives us a better result.

***2) Model 2: Classification Tree***

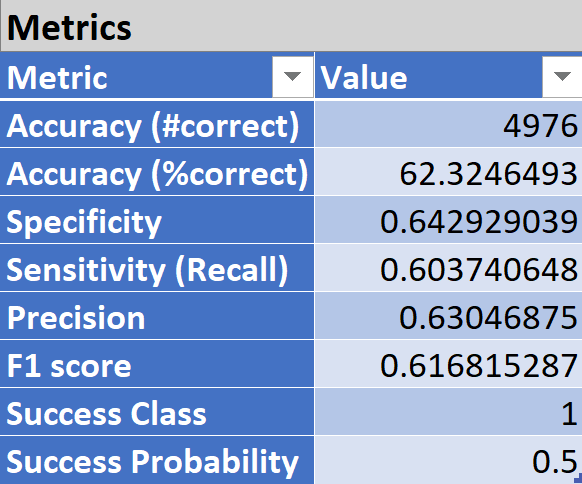
Running classification tree on data, we will create a Full-grown tree and calculate the accuracy of the tree for training, validation, and test data. The data we have has mostly categorical variables and some variables have distinct categories greater than 15, so in order to run the classification model in XLMiner, we have to exclude those categorical variables which have distinct categories more than 15. Below is the output metrics table we got after running the algorithm.

Fig. 3. Classification Tree Error Matrix

The tree we got after running classification tree is shown below. The full-grown tree shown has 99 nodes and 44 decision nodes. Each node in the tree is the drop down tree and classified according to the leaf node it reaches. These newly found classes can be compared to the classification matrix to find their actual membership. The feature importance table we got is shown below, from this, we can see that Local English NameIdentifier has the highest importance.

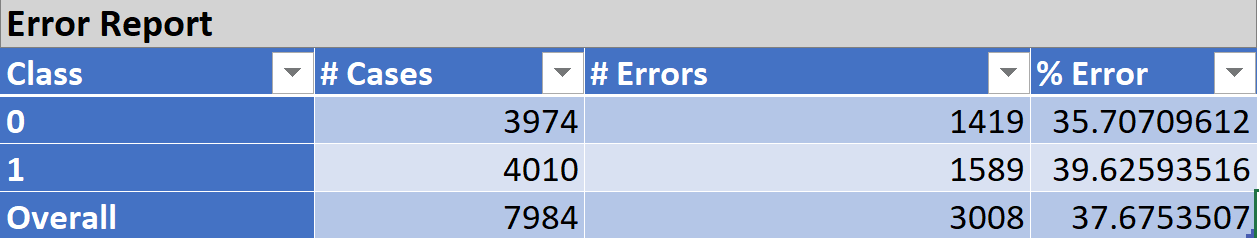


Fig. 4. Classification Tree Data Matrix

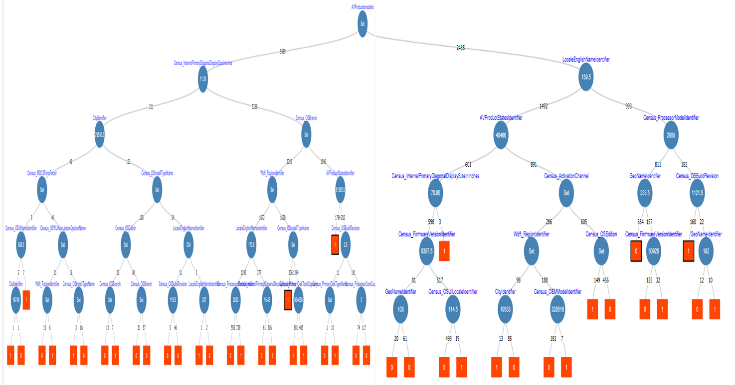


Fig. 5. Classification Tree Data Matrix

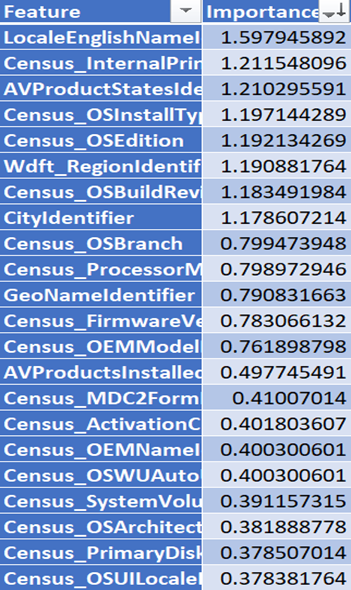


Fig. 6. Classification Tree Feature Importance

***a) Advantages of Classification Tree***

* In the classification tree model, we don’t need to normalize the data, this is one of the biggest advantages.
* In this model there is no need to handle the missing values.
* From a business perspective, the tree is easy to understand, and this can be understood by anyone from business fields without knowing any technical thing.

***b) Disadvantages of Classification Tree***

* In this model, predictors numerical as well as categorical having many values get the preference, which will affect the tree built.
* For better classification, the data should be large, and to create a tree require high computation.

***3) Model 3: Logistic Regression***

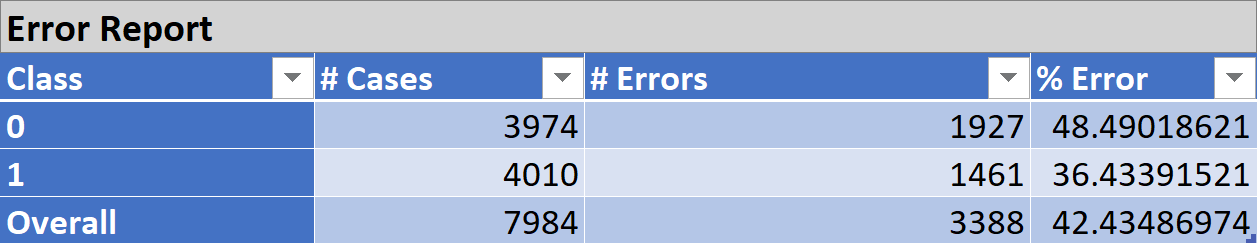
In logistic regression we use the logistic function to output a value ranging from 0 to 1. Based off this probability we assign a class. We can use a confusion matrix to evaluate classification model. Logistic regression is not an algorithm which can be or rather should be used to predict continuous values. Instead, Logistic regression is the algorithm for binary classification where the outcome can be either of 1 and 0 i.e. true or false.

Fig. 7. Logistic Regression Error Metrics

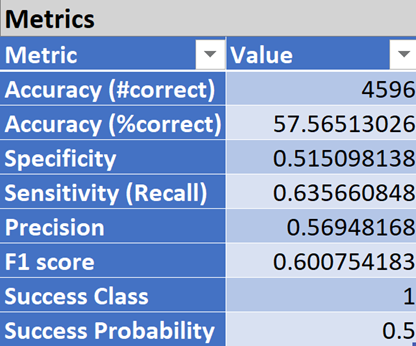


Fig. 8. Logistic Regression Data Metrics

***a) Advantages of Logistic Regression***

* It is efficient and does not require too many computational resources.
* It does not require input features to be scaled as its highly interpretable. Logistic regression does work better if we remove at- tributes that are unrelated to the output variable as well as attributes are correlated to each other.
* Logistic regression is easy to regularize and outputs well calibrated predicted probabilities.

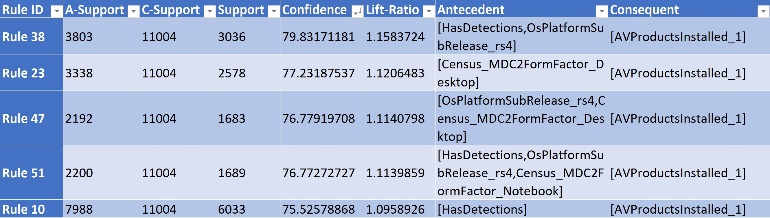
***b) Disadvantages of Logistic Tree***

* In this model, predictors numerical as well as categorical having many values get the preference, which will affect the tree built.
* Logistic regression is high dependence on proper presentation of your data.
* It can be easily outperformed by more complex algorithms as it is not one of the most popular algorithms.
  1. ***Hypothesis 2***

Since gamer devices could have less security, they could be more prone towards getting affected by the malware. Considering the variable \_IsGamer, we can check if the devices used for gaming had malwares detected on them. This will enable us to check the relation between devices used for and not used for gaming. These devices can then be shipped with extra security (either hardware or software) in the form of security chips or software security patches. To prove this hypothesis, we have developed a model using Associa- tion Rules over the columns OsPlatformSubRelease, Wdft\_RegionIden tifier, Wdft\_IsGamer,Census\_MDC2FormFactor,AVProductsInstalled, AVProductsInstalled.

According to the top association rule found in results shown in Fig. 10, we found that if devices is affected with malware (HasDetection) and belongs to rs4 OSRelease platform (OsPlatformSubRelease\_rs4) those devices have 79.83% chance of having antivirus (AVProductInstalled\_1) installed in them.

The top 5 generated association rules with confidence and support are given in the below figure.

Fig. 9. Top 5 Association Rule

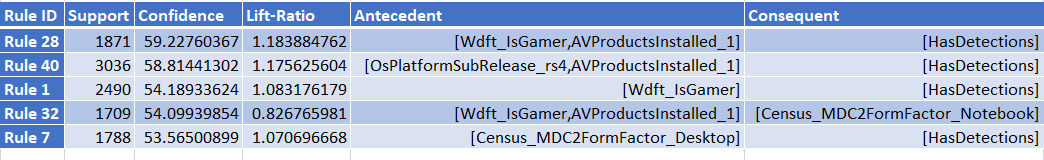


Fig. 10 IsGamer Association Rule

[Wdft\_IsGamer,AVProductInstalled\_1]→ [HasDetections]conf = 59.23%

[Wdft\_IsGamer] → [HasDetections]conf = 54.2%

We also found out that there is 59.23% chance of device being affected with malware attack if it is a gaming device (Wdft\_IsGamer) and has antivirus software (AVProductInstalled\_1) installed in it. And there is 54.2% of chance of gaming device being affected with malware attack.

The above rules show that if the user is gamer and the device have an antivirus product installed, it is still prone to malware attacks.

1. ***Modeling technique’s Advantages:***

In association rule algorithm, association can exist between any it’s attribute which does not happen in standard decision tree algorithm.. A decision tree algorithm will build rules with only a single conclusion, whereas association algorithms attempt to find many rules, each of which may have a different conclusion.

1. **Disadvantages of the modeling technique:**

The disadvantage of association algorithms is that they are trying to find patterns within a potentially very large search space and, hence, can require much more time to run than a decision tree

* 1. ***Hypothesis 3***

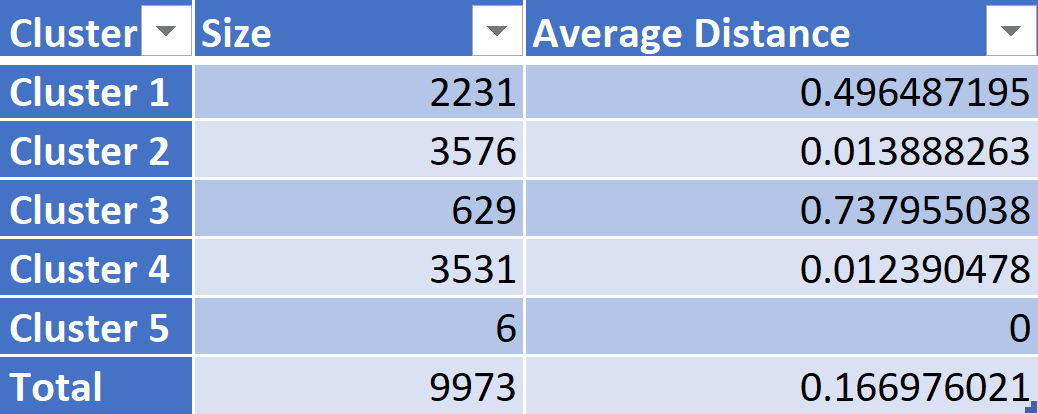
Clustering devices based on malware detection “HasDetection”. According to these clusters, we will find which type of devices are more prone to malware. Based on these clusters it can be seen which devices (Census\_PowerPlatformRoleName) like Desktop, Mobile, Server, Workstation, Slate, ApplicancePC, EnterpriseServer, SOHO server. For this dataset, we have converted the categorical indepen- dent variable to numerical data. Clustering technique used here is k-means clustering with five clusters. Also, for deeper insights, random starts are kept as 3 to find the best cluster.

Fig. 11 Cluster Summary

The Cluster Summary displays the number of records (observations) included in each cluster and the average distance from cluster members to the centre of each cluster. Cluster 3 has the highest average distance of 0.73 and includes 629 records.

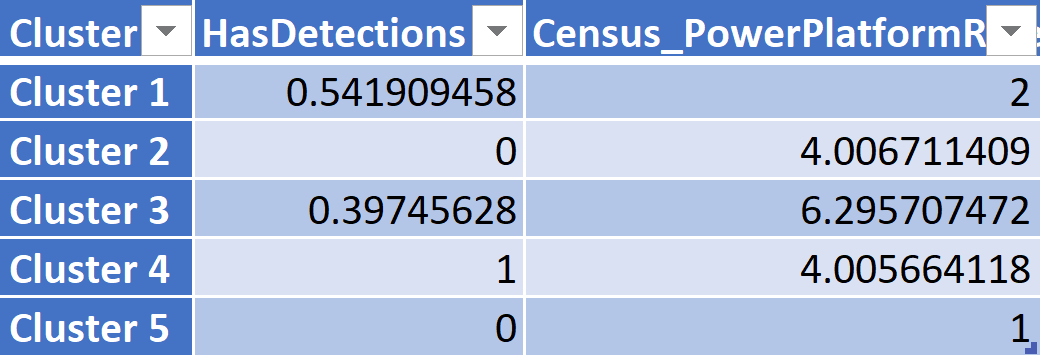
Compare this cluster to Cluster 5, which has the smallest average distance as 0 and includes only 6 members. Based on cluster summary cluster 5 cluster 4 cluster 2 cluster l, cluster 3. Cluster 5 holds ’Mobile’ Device with ’HasDetection’ 1.

Fig. 12 Cluster Centre

The table in Fig. 12 shows the variable values at the Cluster Centres. Cluster 4 has the highest average HasDetections Compare this cluster to Cluster 3, which has the highest average Census\_PowerPlatformRoleName.

Since we have used random start while creating clusters, Best clusters are selected based on sum of squares. Fig. 13 table is the best cluster table with lowest sum of squares values.

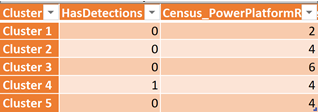


Fig. 13 Best Clusters

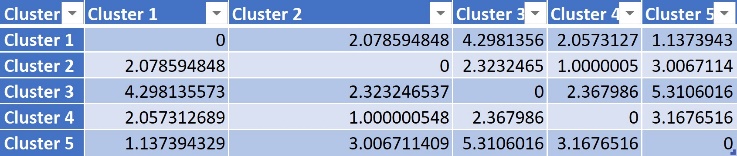


Fig. 14 Best Clusters

From the values in the table of Fig. 14, it is determined that Cluster 1 is very different from Cluster 3 due to the high distance value of 4.298, and Cluster 5 is close to Cluster 1 with a low distance value of 1.137. We cannot perform hierarchical clustering because of XL Miner’s limitation of 1000 rows for hierarchical clustering.

***a) Advantages of k-means clustering:***

* With many variables, K-Means may be computation- ally faster than hierarchical clustering if K-Clustering is small.
* k-Means may produce higher clusters than hierarchical clustering.

1. ***Disadvantages of k-means clustering***:

* Difficult to predict the number of clusters (K-Value).
* Initial seeds have a strong impact on the results.
* The order of the data has an impact on the results.

1. STRATEGIC RECOMMENDATION

After careful investigation and analysis of Microsoft malware data we have come to conclusion that the gaming devices i.e. Desktop, Notebook, Laptop etc are more vulnerable to malware attacks. We have also found many devices which are equipped with basic anti- virus but are still being infected with the different malware attack. Therefore, in future we should device a strategy to equip all the Microsoft devices with strong anti-virus software. Also, we should keep on providing regular security updates to this software.

1. CONCLUSION

After careful investigation, analysis and generating several hypotheses models in XL Miner, we came to following conclusions:

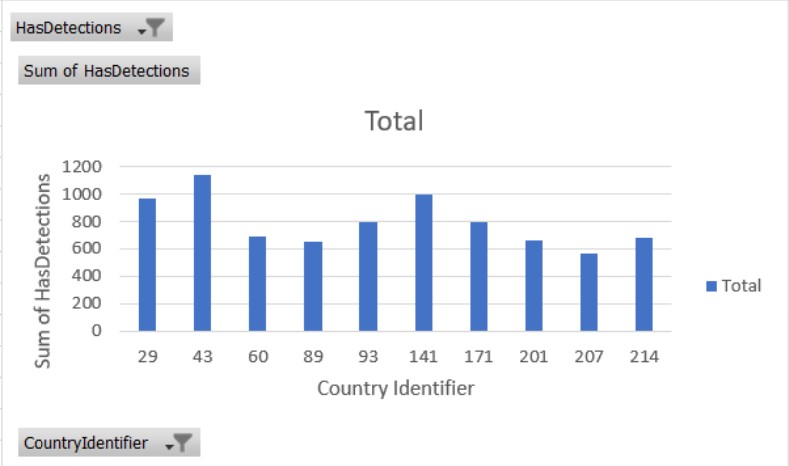
1. We compared various prediction models using same dataset for each model and found out that the KNN gives us the best prediction model with accuracy of 70.29% for the given Malware Detection dataset.
2. If the device is gaming device, it has 54.2% chance of having Malware attack.
3. Looking at these results of hypothesis 3 it can be concluded that Mobile devices are more prone to malware detections.
4. APPENDIX

Fig. 14 Number of malware detection in different countries

The above chart gives us the information about the number of detections of the malware in different countries using countries identifier.

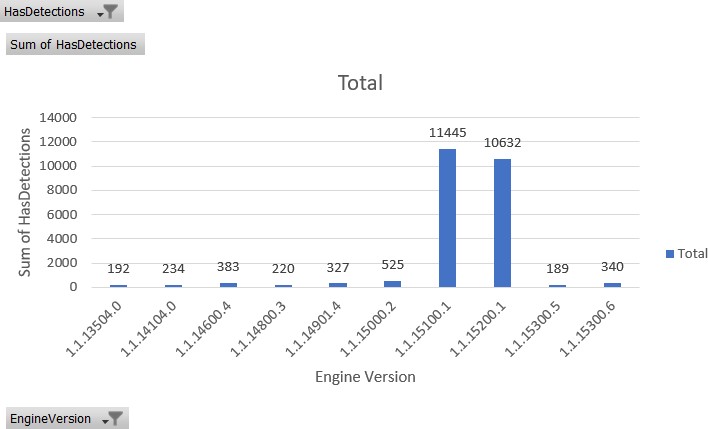


Fig. 15 engine version vs Sum of Has\_detection

The above chart is generated to know which engine versions are more affected with malware, and as we can see that version 1.1.15100.1 and 1.5.15200.1 has the highest count as compared to other versions.

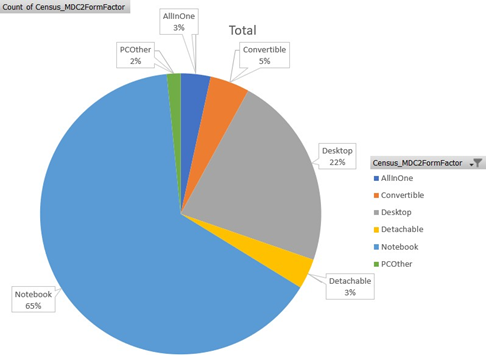


Fig. 16 Census\_MDC2FormFactor Pie Chart

The above pie chart gives insight about the data set distribution with respect to the Census\_MDC2FormFactor. As seen from the below figure the number of notebook machines is 65 percentage and percentage of the desktop is 22 percent, rest other form factors have fewer occurrences in the whole data set.

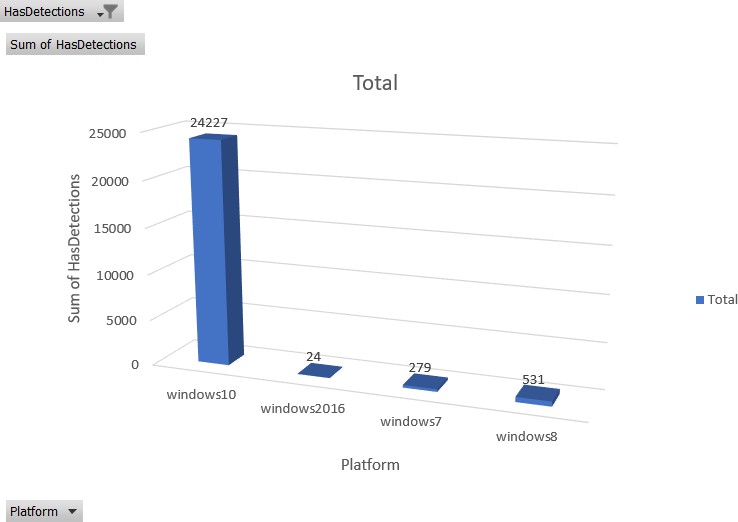


Fig. 17 Census\_OSVersion V/S Sum of HasDetections

The below chart shows the number of malware detection against Census\_OSVersion. For two Census\_OSVersion 10.0.17134.165 and 10.0.17134.228, the number of malware detection is more than 2500 in the whole data set.

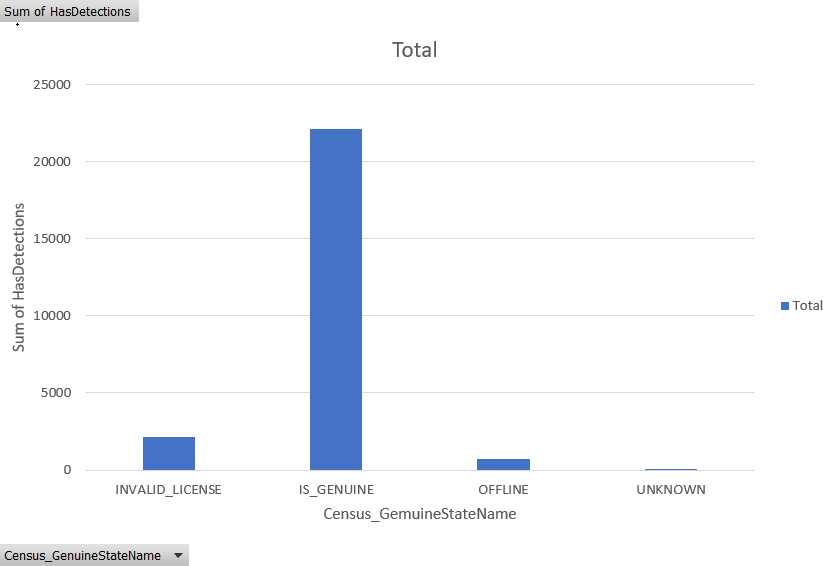


Fig. 18 Census\_GenunineStateName vs HasDetections

This chart in fig. 18 shows the distribution of malware against Census\_GenunineStateName, as visible from the below figure IS\_Genunie state alone contributes more than 20000 rows in the data set.

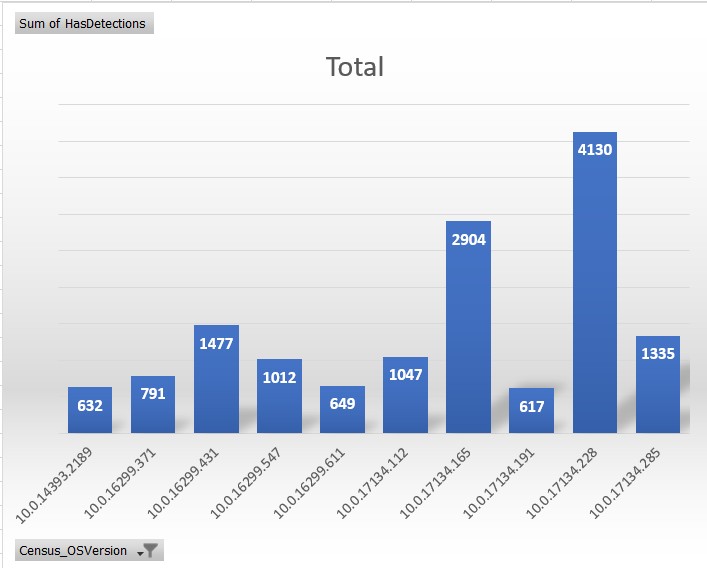


Figure 19: Census\_OSVersion V/S Sum of HasDetections

The above chart shows the number of malware detection against Census\_OSVersion. For two mentioned Census\_OSVersion 10.0.17134.165 and 10.0.17134.228, the number of malware detection is more than 2500 in the whole data set.

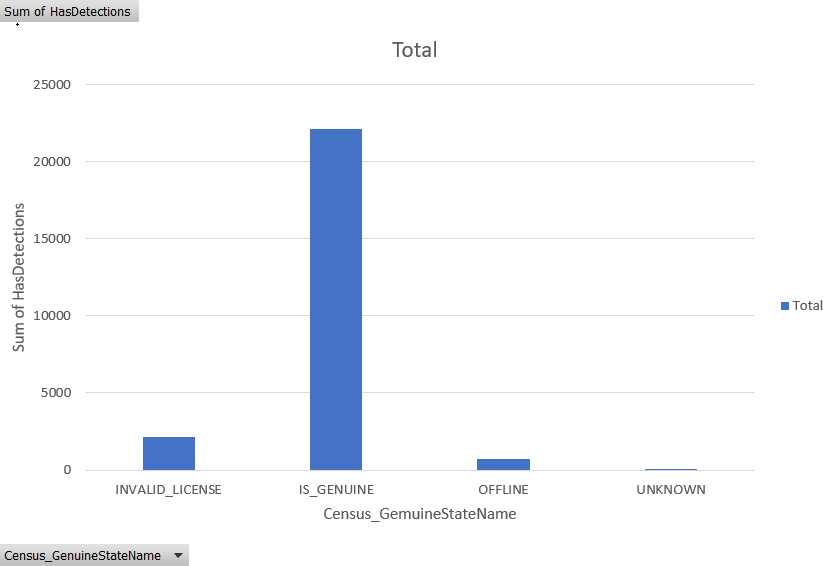


Figure 20: Census\_GenuineStateName V/S Sum of HasDetections

The above chart shows the distribution of malware against Census\_GenunineStateName, as visible from the below figure IS\_Genunie state alone contributes more than 20000 rows in the data set.

1. REFERENCES

[1]<https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-5906-5_838>

[2] <https://www.solver.com/xlminer-capabilities-overview>

[3] <https://en.wikipedia.org/wiki/Cluster_analysis>

[4] G. Shmueli, Peter C. Bruce, and Nitin R. Patel, Data Mining for Business analytics, 3rd Edition, 2016, ISBN: 978-1-118-72927-4

[5]<https://www.kaggle.com/c/microsoft-malware-prediction/data>

[6]<https://help.optimizely.com/Ideate_and_Hypothesize/Design_an_effectiv_hypothesis>

[7]<http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf>

[8]<https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htmDMCON004>

[9] [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov/)

[10] [www.datapine.com](http://www.datapine.com/)

[11] Link.springer.com

[12] ww.thetop10antivirus.com

[13] <https://hbr.org/2013/12analytics-30>

[14] <https://github.com/RubenBranco/Microsoft-Malware-Prediction-2018>