Microsoft Malware Prediction

Rathin Bhargava
IIIT-Bangalore

Ram Srinivasa Bharathy
IIIT-Bangalore

***Abstract*—**
***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The goal of this competition is to predict a Windows machine's probability of getting infected by various families of malware, based on different properties of that machine. The telemetry data containing these properties and the machine infections was generated by combining heartbeat and threat reports collected by Microsoft's endpoint protection solution, Windows Defender.

II. Ease of Use

III. Data Processing and Feature Engineering

A close look at our data reveals that we have 10,000 rows and 83 columns. There were a lot of challenges associated with the data preprocessing like handling null values, categorical data and columns with duplicate data. Here, we show how we dealt with each of these challenges.

A. Data Cleaning

There were some invalid values, like "UNKNOWN", 'Unspecified" in various columns which were converted to nan before any other data preprocessing.

B. Handling columns with null values

We deleted the following columns because more than 30% of the values were null. Imputing these values in any form would just result in biasing the data. Also, it didn't seem like these columns had any relation with the other columns, so we couldn't apply any model to figure out these values based on the other values.

1) DefaultBrowsersIdentifier
2) OrganizationIdentifier
3) PuaMode
4) SmartScreen (Semantically this seems useful, but any imputation here results in a bad accuracy)
5) Census_ProcessorClass
6) Census_InternalBatteryType
7) Census_IsFlightingInternal
8) Census_ThresholdOptIn
9) Census_IsWIMBootEnabled

The following columns were deleted because they didn't convey a lot of information. The values here were either 0 or null.

1) IsBeta
2) AutoSampleOptIn
3) SMode
4) Census_IsFlightsDisabled
5) Census_IsVirtualDevice

On a side note, all the other columns which have been deleted are the ones which have been one hot encoded or have derived features.

The following columns were taken care of by imputing values.

1) RtpStateBitfield
2) AVProductStatesIdentifier
3) AVProductsInstalled
4) AVProductsEnabled
5) IsProtected
6) Firewall
7) UacLuaenable
8) Census_OEMNameIdentifier
9) Census_OEMModelIdentifier
10) Census_ProcessorCoreCount
11) Census_ProcessorManufacturerIdentifier
12) Census_ProcessorModelIdentifier
13) Census_PrimaryDiskTotalCapacity
14) Census_PrimaryDiskTypeName
15) Census_TotalPhysicalRAM
16) Census_InternalBatteryNumberOfCharges
17) Census_OSInstallLanguageIdentifier
18) Census_GenuineStateName
19) Census_FirmwareManufacturerIdentifier
20) Census_FirmwareVersionIdentifier
21) Census_IsAlwaysOnAlwaysConnectedCapable
22) Census_IsVirtualDevice
23) Wdft_IsGamer
24) Wdft_RegionIdentifier

There were a few ways we could impute these values. Imputing with the mean did not make sense, because these values are discrete and specific. They are model numbers, identification numbers, etc. Mode seems to be a good choice for imputing the columns with specific numeric data.

1) Use mode everywhere
2) For Categorical data, impute the null values with -1 and use mode for the numeric data
3) Sample from the inverse probability distribution function. By doing so, the imputed values obey the underlying probability distribution which inherently biases the data the least. This method can be used irrespective of the kind of data - Categorical or numeric.

We have use the third method to impute all the null values. This is done because the column's are roughly independent. PROVE WHYYYYYYYYYYYYY

The column Census_SystemVolumeTotalCapacity was imputed based on Census_PrimaryDiskTotalCapacity. Since there was a high correlation of the system disk being equal to the primary disk (there were no other disks), we made the same assumption for all the values which were missing from the column 'Census_SystemVolumeTotalCapacity'.

C. Handling Categorical data

We have one hot encoded all columns with categorical data where the size of the domain wasn't very high. After a little bit of tweaking, we empirically arrived at the desired domain size, 22. Any more would result in too many columns, any less would lose out on a lot of data. After one hot encoding these columns, the data

The other categorical columns were then label encoded. Label encoding essentially gives every value a random number. There were some columns which while categorical, had a certain order to them. Label encoding these columns would lose the inherent 'order' they possess. Hence, we decided to customize the label encoding for them.

We now explain the customized label encodings, or feature engineering.

1) Versions: All the version numbers were of the form $a.b.c.d$. This format was uniform throughout the five version columns - 'AvSigVersion', 'EngineVersion', 'AppVersion', 'Census_OSVersion' and 'OsVer'. This format essentially means $major.minor[.build[.revision]]$. We made a basic assumption that new update is better than it's predecessors i.e $2.0.1.0$ is better than $2.0.0.999$. To establish such an ordering, we converted the string into a number by multiplying the major, minor, build and revision numbers with the appropriate powers of 10 and adding them up.

D. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections IV-A–IV-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BIBTEX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BIBTEX to produce a bibliography you must send the .bib files.

LATEX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LATEX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [**?**].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [**?**]. The sentence punctuation follows the bracket [**?**]. Refer simply to the reference number, as in [**?**]—do not use “Ref. [**?**]” or “reference [**?**]” except at the beginning of a sentence: “Reference [**?**] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [**?**]. Papers that have been accepted for publication should be cited as “in press” [**?**]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.