

# **THE POST-HOC EXPLAINABILITY OF A BLACK-BOX MALWARE DETECTION MODEL USING TABULAR AND NLP- BASED METHODS**

Presentation by Junel Alje Isanan



# ABSTRACT

**Objective: Establish user trust via XAI**

## **Key Points**

- **An exponential surge of malware from 100M to 700M types (2012-2018). (Saxe, 2018)**
- **Transition to ML for automated detection. Identification of malware through digital signature is cumbersome (Saxe, 2018)**
- **Emphasis on deep learning for evolving threats.**

# ABSTRACT

**Objective: Establish user trust via XAI**

## **Key Points**

- **Deep learning models achieve high accuracy in detection, but lack transparency.**
- **Use XAI tools to uncover influential features in malware detection.**
- **Goal is to enhance transparency, foster trust, and improve detection methods.**

# INTRO

## Background & Problem:

**Malware threatens systems; projected ransomware damages:  
\$250B by 2031. (Cybercrime Magazine, 2023)**



# INTRO

## Objectives

- Source a malware dataset
- Train machine learning models from the data
- Use XAI tools to determine which features influenced the prediction
- Produce feature visualizations of individual predictions

# **SIGNIFICANCE**

## **Significance of the Study**

- **Malware remains a persistent threat in cyberspace.**
- **Black-box models enable rapid identification of new malware strains / detection of zero-day malware.**
- **Despite high accuracy of these models, explainability is crucial for user trust.**

# **SIGNIFICANCE**

## **Significance of the Study**

- **This research addresses the need for explainability in malware detection, similar to its importance in finance and healthcare.**
- **The study contributes to building trust in AI-driven malware detection, improving security decisions, and potentially saving lives and resources.**




# LITERATURE REVIEW

Various types of malware with specific functions include:

- Adware or spam
  - Viruses
  - Worms
  - Trojans
- Ransomware

Which steal, extort, or destroy data, or copy themselves into the host machine








# LITERATURE REVIEW

**Commercial antiviruses have limited capability to detect new malware due to their reliance on databases of digital signatures, they have to update their software frequently.**


**The research's direction would be to test out-of-sample malware data.**






# LITERATURE REVIEW

**XAI (Explainable AI) - refers to techniques and methods that make artificial intelligence (AI) models' decisions and workings understandable to humans.**

- **Builds trust and transparency in AI**
  - **Enables better decision-making**
  - **Ensures fairness and accountability**
- 



# **MATERIALS & METHODS**

## **Host Machine:**

**OS: Windows 10 Pro 64-bit.**

**Processor: Intel Core i3-10100F, 3.6 GHz, 8 cores**

**RAM: 16GB RAM.**





# **MATERIALS & METHODS**

**Development Tools:  
Jupyter, Git, Python.**



# **MATERIALS & METHODS**

**Data Cleaning - Transforming data to become usable,  
standardized**

**Data preprocessing - Machine-readable data**






# **MATERIALS & METHODS**

**Model Training - Feeds the data into the black-box model**

**XAI Processing - Feeds the data and model into an XAI Model for individual prediction.**





# **MATERIALS & METHODS**

**Dataset Used:**

**Ember Dataset (train\_features\_5.jsonl)**

**Used only the imports, section, and string information.**





# **MATERIALS & METHODS**

**For headers, obfuscated and packed headers were recorded.**

**For imports, only the top 500 most frequent functions were considered (due to dimensionality issues)**



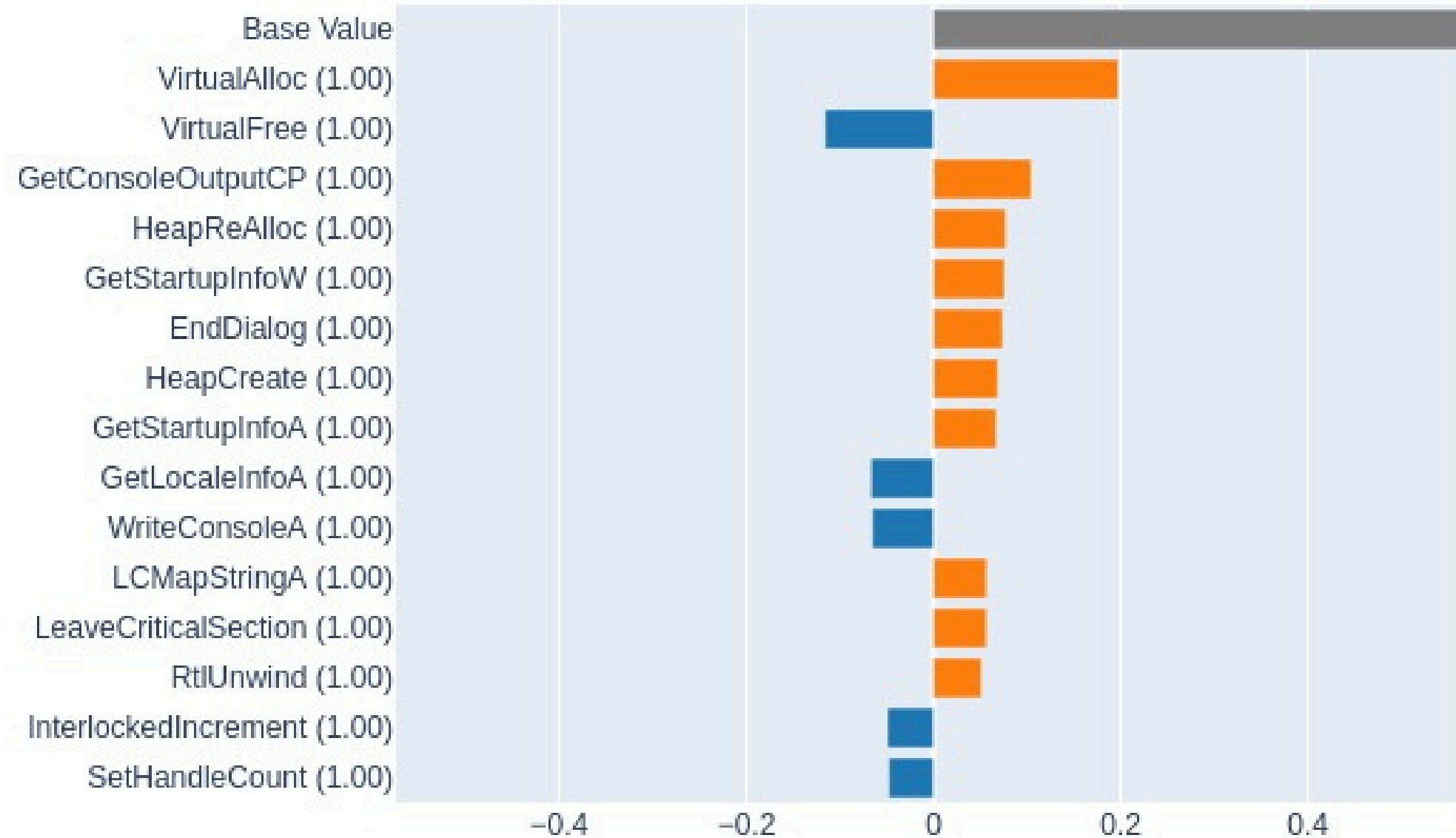


# MATERIALS & METHODS

'n86sq', '.v51h', '.dr45', 'LvFS9LeA', '.late', '.dart.', '.star', '.y0da', 'xzqltxgu', 'ygwnpamx',  
'ebsll', 'vscorcj', 'PEPO', 'PEP1', 'vdbrsyg', '.9h5o', '.8dne1i', 'Rmz%op2|', 'Hey', 'you!!',  
'PLeaSe', 'SToP!', 'FuCK', 'youRSeLF', 'STuPiD', 'Guy!', 'ebgrawn', 'cuopnlh', 'vrzaoyh',  
're9o7i', '.y8r8i', 'BISJIJSH', '.upx10', '.upx11', '.newdata', '.export', '.annoter', 'ORNO',  
'ORN1', '.garr', '.gaa8', '.ddesw', '.grdn9', '.asdw', '.qwec', '.3kga', '.el8dtc', 'ObAR9RE1',  
'w', '.reloc ', 'zbknwnp', '.network', '.zpxc', '.cdnr5a', '.01r6', '.uaimo', '.3yfid', 'FFFF',  
'wotir', '6269hzje', '6z3wwfsj', '639yobka', 'shyjgvzm', 'bccbvicg', '.etrtrd', '.w6t9',  
'u9o1m', '.l9ew1', 'jpuwkqv', '.4is6eh', '.5hsa', '.oe6d', '.DDDD', '.9dsoty', '.c3tt', '.r529',  
'etd7', '.72r48', '.goheu', '.l9eb', '.1a7e09', 'pd\_rec0', '.zexe', '.mapodat', 'gbesxmm',  
'nsh0', '.nsh1', 'selinuv', 'ffshrbrm', '.o0wi', '.yeieh', '.n6fgru', '.i5sg', '.togr55', 'camel131',  
'rustc', 'a%ti);!['', 'oifdutmx', 'kmkndxfx', '.TCSHARE', 'gqqstuu', 'idudxls', 'kollej', 'pee',

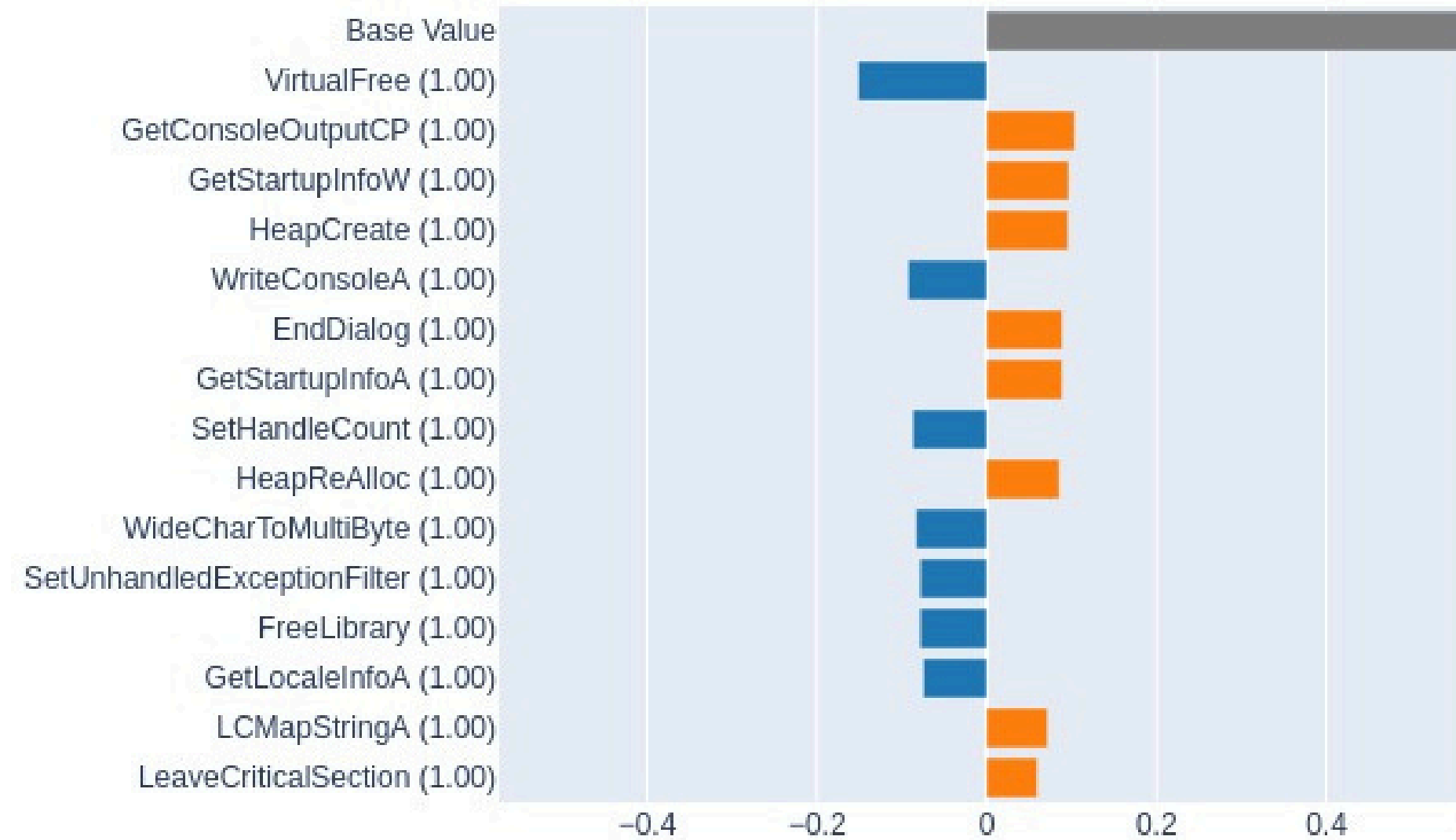
# RESULTS

Actual: 1 | Predicted: 0.926



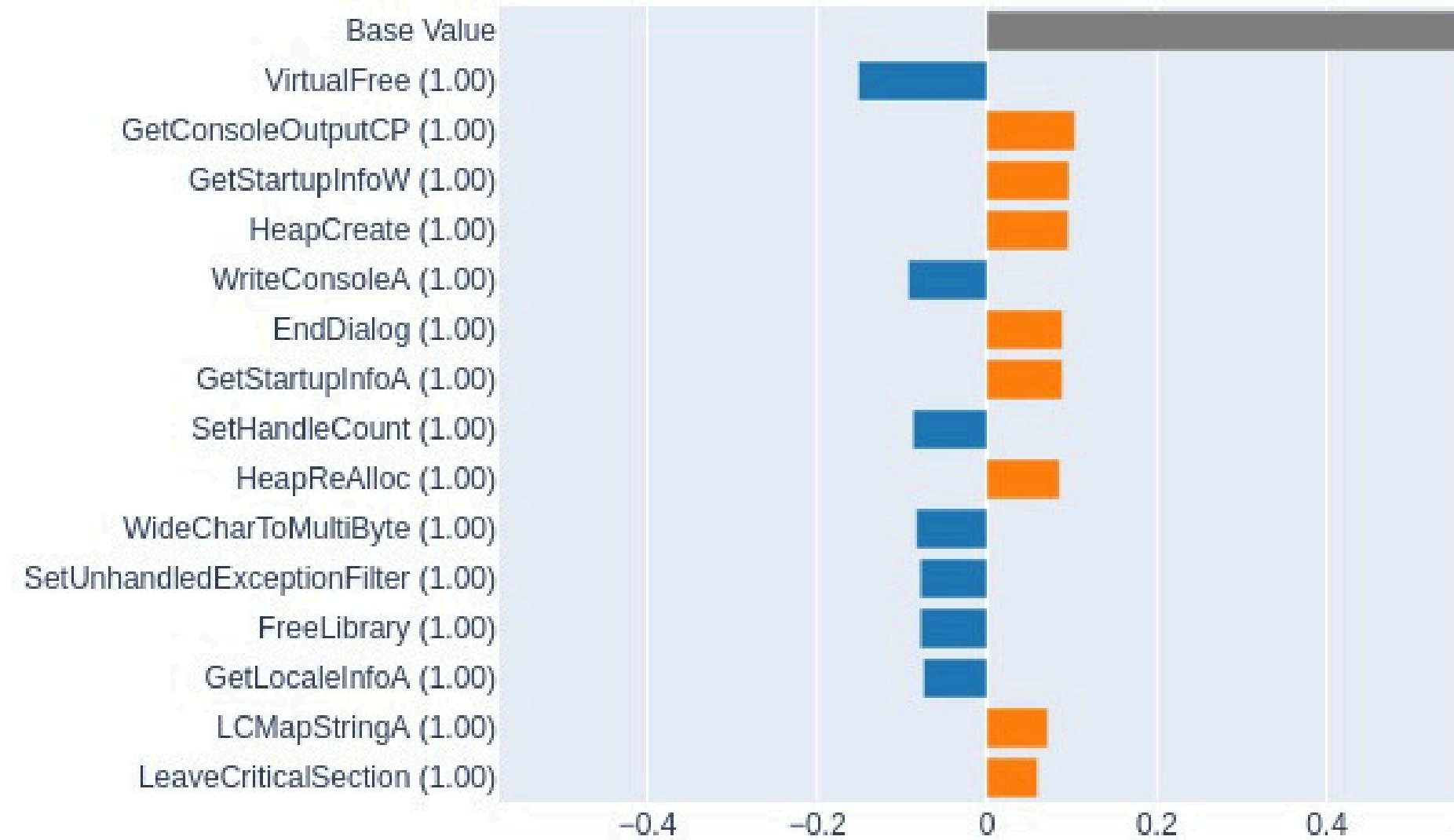
# RESULTS

Actual: 1 | Predicted: 0.406



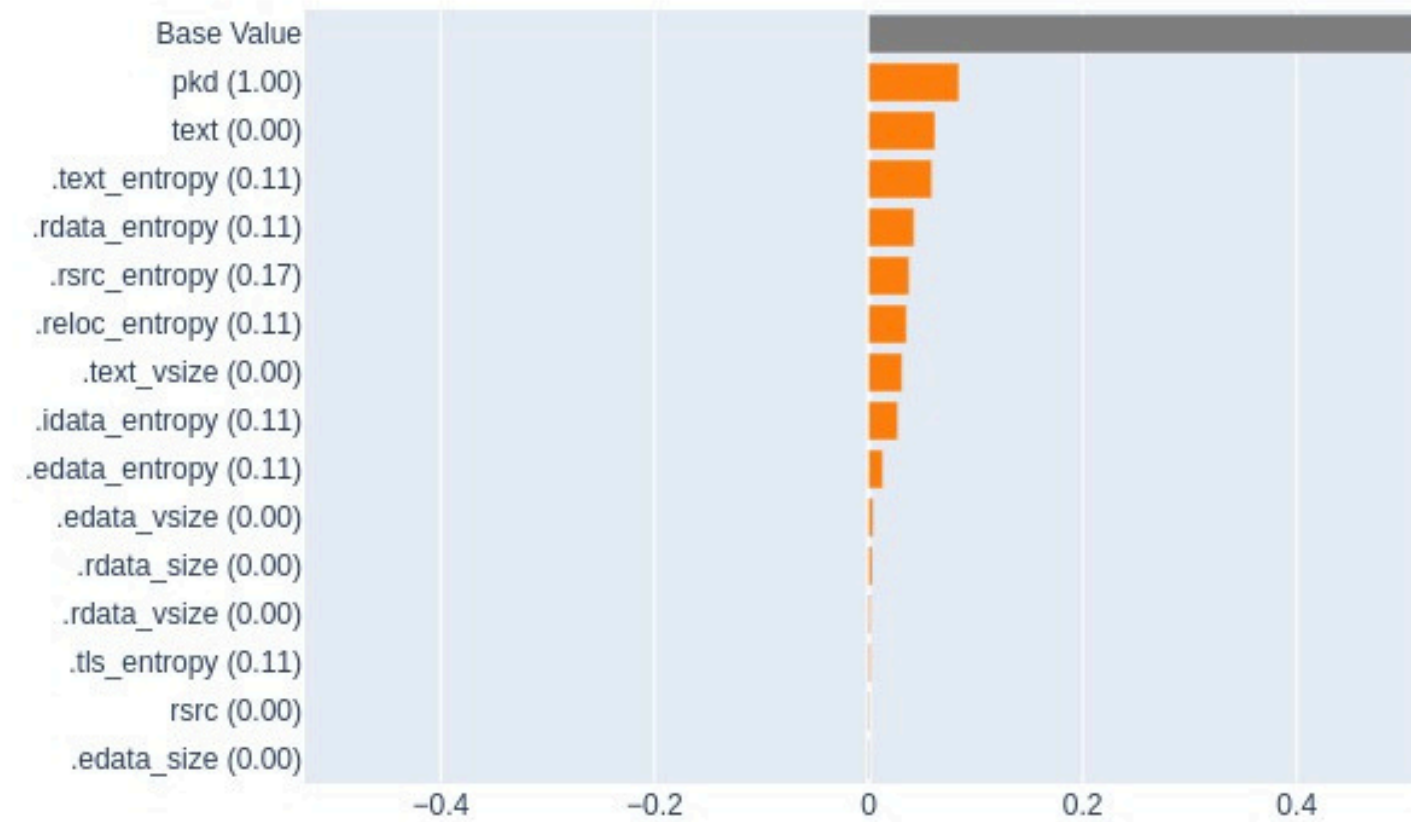
# RESULTS

Actual: 1 | Predicted: 0.406



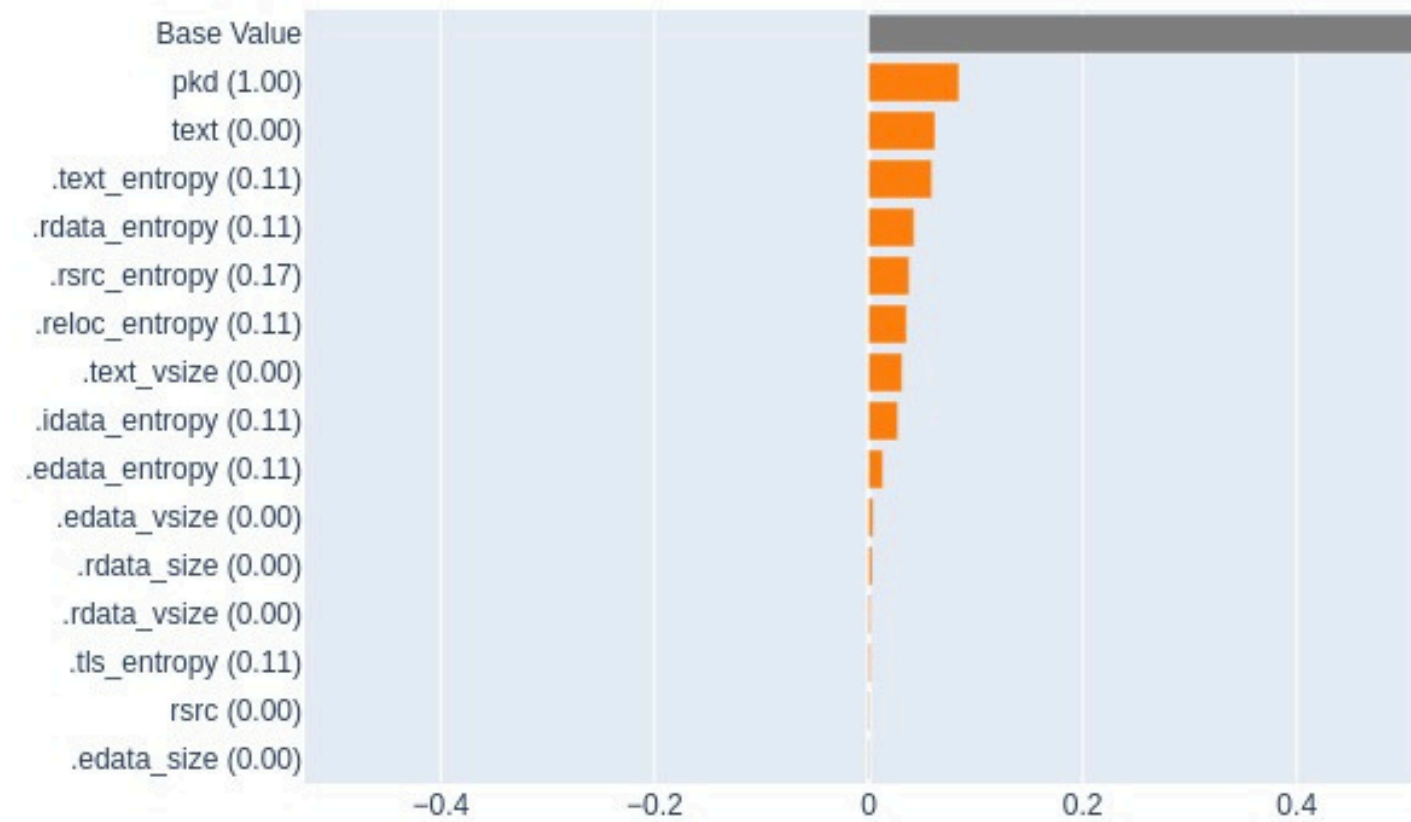
# RESULTS

Actual: 1 | Predicted: 0.934



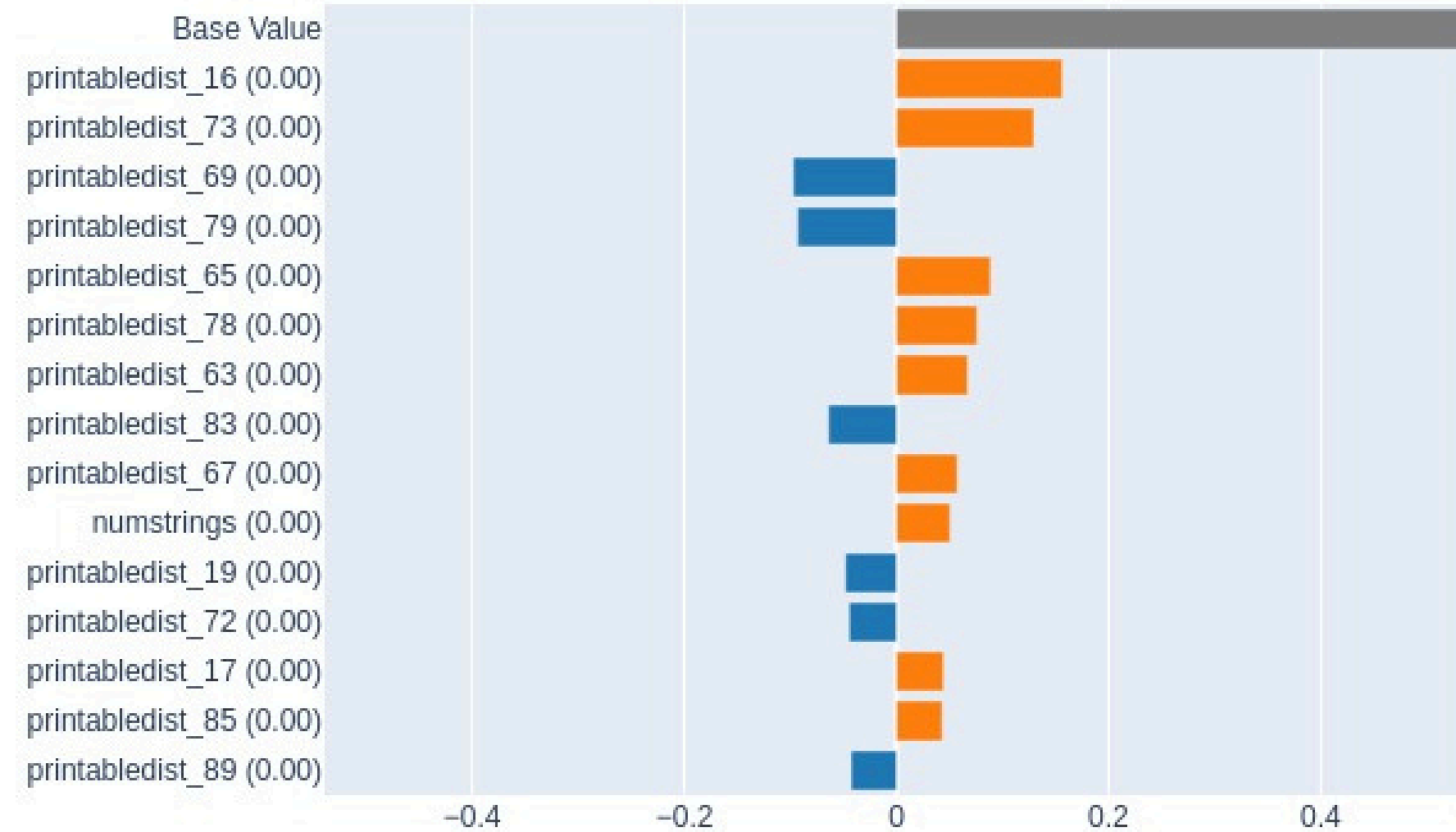
# RESULTS

Actual: 1 | Predicted: 0.934



# RESULTS

Actual: 1 | Predicted: 0.973





## REFERENCES

Saxe, J., & Sanders, H. (2018). Malware data science: Attack detection and attribution. No Starch Press.

Sausaulito, C. (2023, July 10). Global ransomware damage costs predicted to exceed \$265 billion by 2031. Cybercrime Magazine. <https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-reach-250-billion-usd-by-2031/>





# THANK YOU

Presentation by Junel Alje Isanan

