College of Computers and Information Technology

Department of Computer Engineering

Advanced computer networks

B2B Technographic Data

Joud Fawaz Alosaimi – 44104133

Supervised by :

Nada Al-Tuwairiqi

Department of Computer Engineering

College Computers and Information Technology

Taif University, KSA

Introduction

The **Business Technographic Dataset for Saudi Arabia**, provided by Techsalerator, offers a comprehensive and analytically rich resource tailored for businesses, technology providers, and market researchers seeking to gain insights into the technological landscape of companies operating within the Kingdom. This dataset systematically captures detailed information on technology stacks, digital tools, and IT infrastructure in use across various sectors. In the context of this study, the dataset is employed for a **classification task**, where the objective is to predict whether a given technology is deployed **behind a firewall** based on company and technology-related attributes. The dataset (Kaggle)

Data description

This dataset includes seven columns, six of them are feature variables, and the last one is the target, labeled as "Behind Firewall". Each row provides details about a specific technology observed within a company. The features capture various aspects such as the company domain, the technology in use, and when it was detected. Below is a closer look at what each column represents:

- **Website Domain** (Categorical (string)): Identifies which company the observed technology belongs to. Useful for grouping technologies by company.
- **Ticker** (Categorical (string)): The financial market ticker symbol of the company, it's non-essential for classification.
- **Technology Name** (Categorical (string)): A primary feature for determining how advanced or secure the company's tech environment is.
- **Technology ID** (Categorical (UUID)): A unique identifier for each technology tool or platform.
- **First Seen At** (Date): Used to calculate usage duration or to track technology adoption over time.
- **Last Seen At** (Date): with First Seen At, you can derive the usage duration (in days), which may be a strong predictor of system criticality or firewall usage.
- **Behind Firewall** (Boolean (True/False)): Indicates whether the detected technology is used behind a firewall (internally hosted or secured) or not. ○ **True** → Technology is deployed internally and protected behind a firewall. ○ **False** → Technology is accessible publicly or hosted externally.

Data preprocessing

The dataset consists of **308 samples**, which makes it relatively small and manageable. As a first step in preprocessing, we removed the **"Ticker"** column since it contained many missing values and did not contribute useful information for predicting the target variable (Behind Firewall).

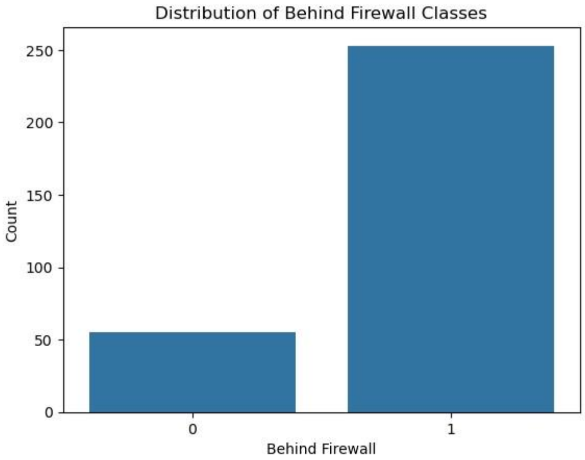
Next, we applied **label encoding** to convert all categorical string attributes into numerical format, allowing them to be processed effectively by the machine learning models.

We then split the dataset into **75% for training** and **25% for testing**. This split ensures that the model has access to enough data (231 samples) to learn meaningful patterns, while still keeping aside a reliable portion (77 samples) for testing its ability to generalize to new, unseen data. This balance helps reduce the risk of overfitting and provides a fair evaluation of model performance.

Finally, we applied **standardization** using "StandardScaler" to scale the feature values. This step is particularly important for models that are sensitive to feature magnitudes, such as KNN, SVM, and ANN.

Data visualization

Count Plot: A bar chart used to display the frequency of observations in each category. In our case, we used it to visualize the distribution of the target classes 0 and 1. The plot clearly shows a significant imbalance, with class 1 appearing far more frequently than class 0.



Models' performance Analysis

Name	Decision tree	Random forest	KNN model	SVM model	Naivebayes	ANN model	Logistic regression
Accuracy	79 %	94 %	91 %	92 %	91 %	92 %	88 %
Precision	93 %	93 %	92 %	92 %	92 %	93 %	89 %
Recall	82 %	100 %	98 %	100 %	98 %	98 %	98 %
F1-score	87 %	96 %	95 %	96 %	95 %	96 %	94 %

The performance when the class = 1 (protected behind a firewall)

Looking at the table, the Random Forest model demonstrated the best overall performance across all metrics. It achieved 94% accuracy, 93% precision, 100% recall, and a strong 96% F1-score. This performance can be attributed to the ensemble nature of Random Forest, which leverages multiple decision trees to reduce overfitting and enhance generalization.

The Decision Tree model, on the other hand, had the lowest overall performance, with an accuracy of 79% and an F1-score of 87%. Although it showed a high precision of 93%, its recall was only 82%, indicating that it struggled to correctly identify all relevant instances of the positive class — likely due to its sensitivity to data splits and tendency to overfit.

Models like KNN, SVM, Naive Bayes, and ANN performed consistently well, each reaching accuracy levels above 91%, with F1-scores of 95% or higher. Their balanced precision and recall values (around 92–98%) indicate that they were effective at both detecting and correctly classifying instances. Notably, SVM achieved 100% recall, meaning it did not miss any positive cases.

Logistic Regression, while slightly behind the others in accuracy (88%), still delivered a respectable precision of 89%, recall of 98%, and F1-score of 94% — showing that even simpler linear models can provide strong results when the data is well-preprocessed. In summary, most models performed well in this classification task, particularly those capable of capturing complex patterns such as Random Forest, ANN, and SVM. The best model overall was Random Forest, and the weakest in terms of combined performance was Decision Tree, despite its strong precision.

Conclusion

I picked this dataset because it focuses on the Saudi Arabian technology landscape, which is both under-explored and highly relevant in the modern digital economy. The dataset provides real-world business technology data, making it valuable for analysis. Its features reflect practical situations companies face, such as firewall usage decisions. Understanding which technologies are used behind firewalls can guide cybersecurity strategies. Such insights are crucial for IT providers and decision-makers in the region. The classification task is also straightforward, making it suitable for model comparison. Additionally, the data’s structure allows for clear preprocessing and transformation steps. Working with it helped me practice encoding, scaling, and model evaluation techniques. The best-performing model was Random Forest, with 94% accuracy and 100% recall. This model outperformed others due to its robustness and ability to handle complex patterns. It minimized overfitting while delivering consistent results across metrics. This shows how ensemble learning is effective with small, imbalanced datasets. One key insight is that usage duration may correlate with security decisions. Technologies used longer were more likely to be behind a firewall. This implies companies secure long-term tools more than recently adopted ones. The model learned this pattern efficiently, reinforcing the value of good feature design. I also observed that simple models like Logistic Regression still performed quite well. In conclusion, this dataset helped me link technical work to real-world cybersecurity insights. It was a meaningful and informative learning experience overall.

GitHub Link

https://github.com/Joudfz/B2B-Technographic-Data/tree/main/B2B%20Technographic%20Data-JoudAlosaimi-44104133