Transport and Telecommunication Institute

Faculty of Engineering Science

Machine Learning and Predictive Analytics

Interpretable Models for Detecting Suspicious Users on Social Media

Student: Gonzalo, Gamez

Student's ID: st83446

Study Group: 4303MDA

2024

Abstract

In 2023, there were 4.9 billion social media users, with a projected increase to 6.05 billion by 2028 (Dixon, 2024b). A subset of these users create detrimental content, including fake news, hate speech, and rumors (Gongane, Munot, and Anuse, 2022). While machine learning is often used to detect detrimental content, this research focuses on identifying suspicious users based on features of the user and the features of content published by the user. This approach aims to complement existing content-based moderation strategies. To foster user trust, the models developed should be able to provide explanations for identifying suspicious users. This study employs Logistic Regression and Decision Tree models, prioritizing interpretability—a trade-off for potentially higher accuracy achieved by more complex models. Using a TikTok user dataset, the models were optimized to identify users under scrutiny or banned, evaluate the accuracy of suspicious user detection, and interpret the decision-making process of the best-performing model. The final Decision Tree model, after addressing data imbalances and model adjustment, achieved an F1 score of 0.47, using features like video view count, presence of specific keywords, and video duration.

Repo:https://github.com/Takosaga/fall_24/blob/main/machine_learning_and_predictive_analytics/detecting_suspicious_tiktok_content_creators/notebooks/report.ipynb

Word Count: 1954

Introduction

As of January 2024 (Dixon, 2024a), there were 1.582 billion monthly active users on the social media platform of Tiktok. TikTok employs a combined approach of automated detection and human review to enforce its community guidelines (TikTok, 2024). The moderation process begins with automated systems that scan content for potential violations. These systems include algorithms trained to identify various forms of content that violate guidelines or terms of service of the platform. Content flagged by these automated systems is then subject to review by human moderators. TikTok recognizes content may not be detected by this method so it encourages reporting in-app and on the website by other users along with review for content that has gained popularity. However, current moderation practices primarily focus on content itself, rather than the patterns of behavior exhibited by users who create harmful content. This research seeks to address the gap by investigating the potential of machine learning to identify suspicious user behavior as a complementary approach to existing content based moderation strategies.

This research acknowledges the trade-off between model accuracy and interpretability as illustrated in figure 1 (Shi et al., 2020). While more accurate models like neural networks may offer higher accuracy, their complexity makes it difficult to interpret their decision-making process. Therefore, this study focuses on interpretable models like logistic regression and decision trees, prioritizing interpretability, even if it means potentially sacrificing some degree of predictive accuracy. This prioritization is justified by the critical role of trust in AI systems, as shown in the research by Afroogh et al. (2024). This source emphasizes that trust is not merely a desirable outcome but a fundamental requirement for the development, design, and deployment of AI technologies. By using interpretable models, this research aims to not only identify suspicious user behavior but also to provide clear explanations for these classification.

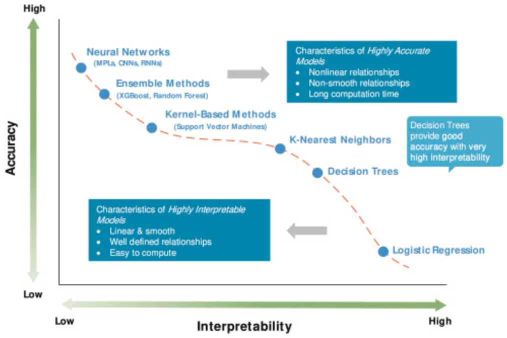

High
Accuracy
Low
Neural Networks
(MPLs, CNNs, RNNs)
Ensemble Methods
(XGBoost, Random Forest)
Kernel-Based Methods
(Support Vector Machines)
K-Nearest Neighbors
Decision Trees
Logistic Regression
Characteristics of *Highly Accurate Models*
• Nonlinear relationships
• Non-smooth relationships
• Long computation time
Decision Trees provide good accuracy with very high interpretability
Characteristics of *Highly Interpretable Models*
• Linear & smooth
• Well defined relationships
• Easy to compute
Low
Interpretability
High

Figure 1. The trade-off relationship between model accuracy and interpretability (Shi et al., 2020)

This research will deploy Logistic Regressions and Decision Trees on a dataset of TikTok users which includes a users ban status to:

i. Optimizing model to Identify the user who created video is suspicious if they are under scrutiny or banned

ii. Evaluate accuracy of detecting suspicious users

iii. Interpret the best performing model

Data Description and Preparation

A dataset provided from kaggle (Shokirov, 2023) will be analyzed. The dataset contains users and the video published by them on TikTok which includes 19,383 entries with 12 attributes. Attributes include:

- Ban status of user who published video: active, under scrutiny, or banned
- Verified status of user who published video: verified or not verified
- Numerical data of published video which are views, likes, shares, downloads, duration in seconds, and comments
- Claim status of published video which has been identified as an opinion or a claim.
- Transcription text of published video

298 entries were found to have missing values which were dropped. According to figure 2, we found outliers for likes, comments, shares and downloads. Looking at figure 3, views, likes, comments, shares and downloads have right tail distributions.

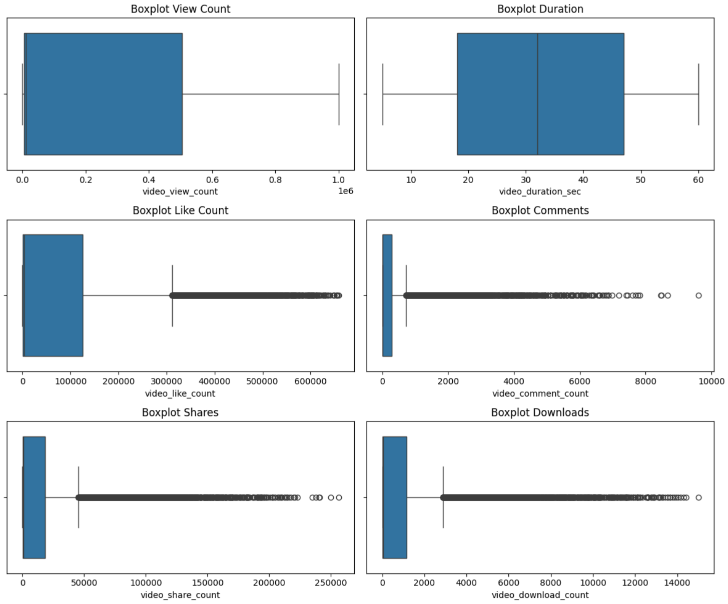


Figure 2. Boxplots of numerical values

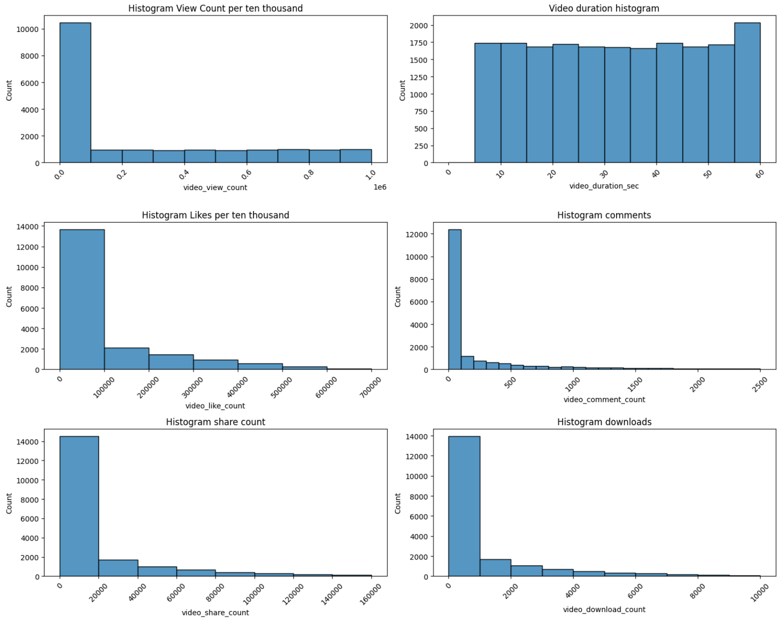
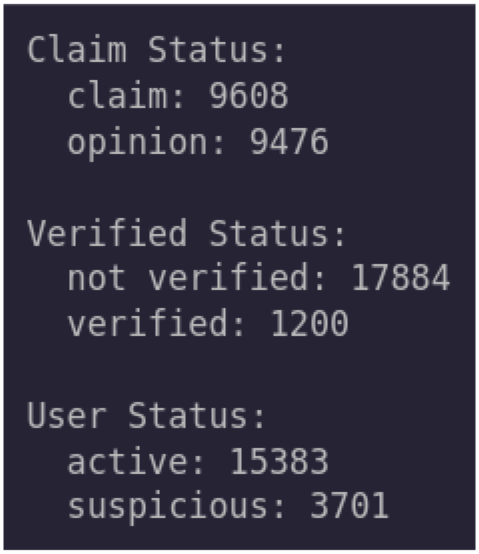


Figure 3. Histograms of numerical values

Categorical data of verified and claim status were encoded by scikit-learn's `OneHotEncoder` to be processed along with the first category to be dropped from each feature during encoding to avoid multicollinearity. Ban status was altered to user status to where a suspicious user is considered to have a status of under scrutiny or banned. User status is the target variable and was also transformed into a numerical representation by scikit-learn's `LabelEncoder`. Figure 4 shows that there is an imbalance of values for user status.



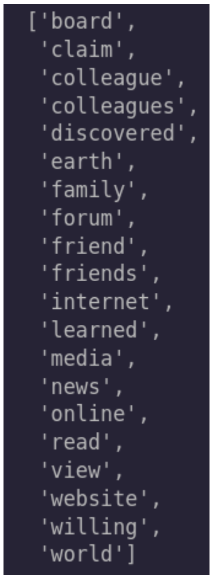
```
Claim Status:
  claim: 9608
  opinion: 9476

Verified Status:
  not verified: 17884
  verified: 1200

User Status:
  active: 15383
  suspicious: 3701
```

Figure 4. Amount of categorical data

Text was decided to be encoded by Bag-of-Words method with the following configurations: tokenization and lowercasing, removal of English stop words , and a vocabulary limited to the top 20 most frequent terms which are shown in figure 5. Bag-of-words was chosen since it can be quickly implemented along with other scikit-learn's packages and is easier to interpret results as opposed to TF-IDF. This resulted in a sparse matrix representation where each row represented a document and each column represented the count of a specific word.



```
['board',
 'claim',
 'colleague',
 'colleagues',
 'discovered',
 'earth',
 'family',
 'forum',
 'friend',
 'friends',
 'internet',
 'learned',
 'media',
 'news',
 'online',
 'read',
 'view',
 'website',
 'willing',
 'world']
```

Figure 5. Top 20 most frequent terms

Before training any machine learning models, the data was split into training and testing sets to evaluate model performance on unseen data and prevent overfitting. The data was split into training and testing sets using scikit-learn's `train_test_split` . The feature matrix (`X`) and the target variable (`y`) were divided, with 20% of the data allocated to the test set and the remaining

80% to the training set. To ensure reproducibility and consistent results, a random state of 42 was used.

Model Development and Improvements

This research uses Logistic Regression and Decision Trees as baseline models due to their inherent interpretability.

- *Logistic Regression:* A linear model suitable for binary classification problems, which provides insight into the influence of each feature of the prediction by examining feature coefficients
- *Decision Tree:* A non-linear model that splits data based on feature values, forming a tree-like structure. Its visual representation makes it easy to interpret the decision-making process, ensuring interpretability

An initial model training was run to make a baseline reference where only encoding and standard scaling to numerical data was applied.

Each model requires different adjustments to improve results. Logistic Regressions are required to handle outliers and multicollinearity. Decision Trees require pruning to prevent overfitting. Both models require resampling since the target has an imbalance in the dataset.

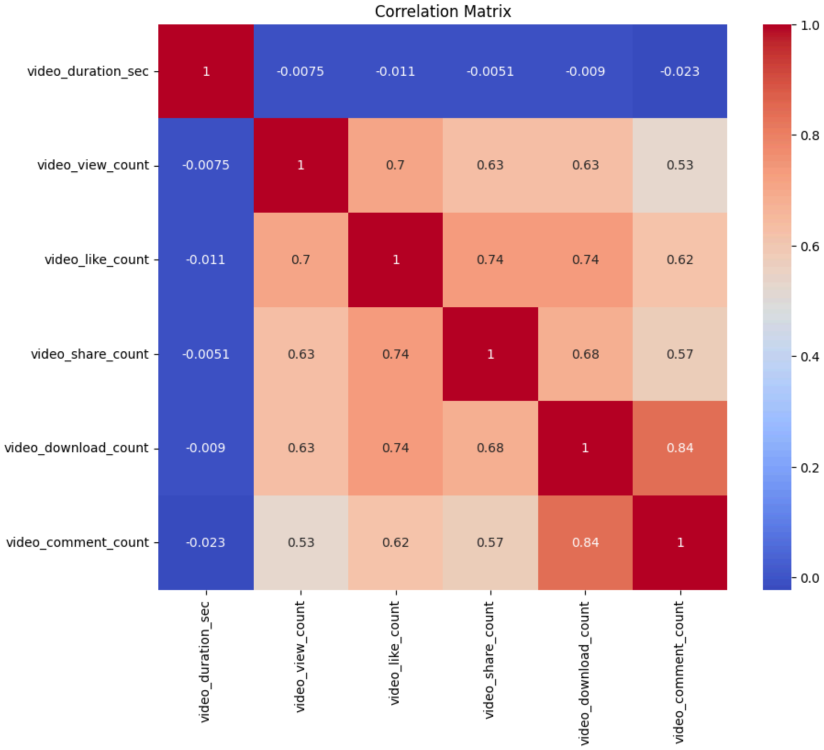


Figure 6. Heatmap of numerical values after outliers removed

For the logistic regression model, the outliers were detected by interquartile range (IQR) and then were removed if they laid outside. Figure 6, shows that downloads and comments were highly correlated with others features having medium or no correlation. Download count was dropped to avoid multicollinearity.

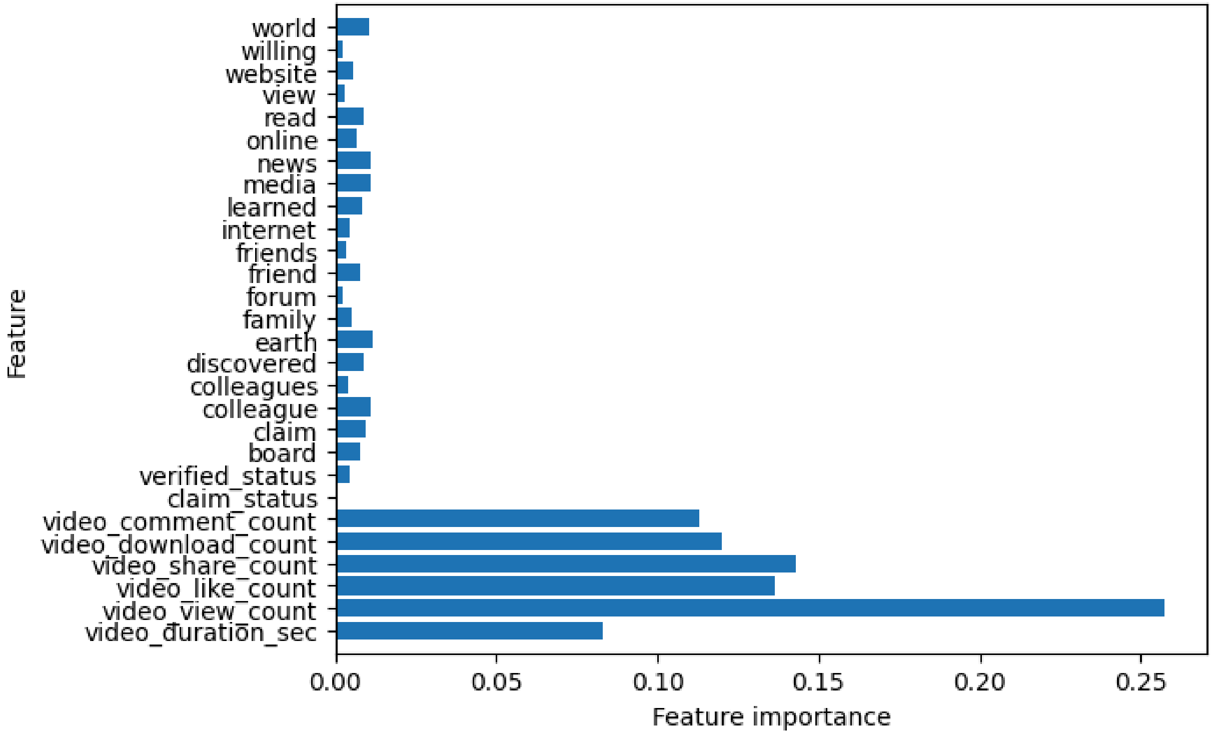


Figure 7. Importance of features for decision tree

According to Figure 7, there are 6 features that have over .05 importance for the Decision Tree. Model could be simplified by removing less important features, only claim status was dropped since it was the one that showed no significance in importance. Though feature importance does not indicate max depth for the model, we will use max depth of 5 and 6 to start.

Resampling is required since Figure 4 shows the imbalance of target values for both datasets with a decision made to use a technique of over-sampling, under-sampling and a combination of both chosen from the imbalanced-learn (2016) package. For over-sampling, the Synthetic Minority Oversampling Technique (SMOTE) generates synthetic examples for minority class samples by interpolating between existing minority samples. With random undersampling class balance is made by randomly removing samples from the majority class. SMOTEEN (SMOTE + (ENN)Edited Nearest Neighbors) is a hybrid method combining SMOTE which generates samples while ENN removes noisy samples.

Model Evaluation

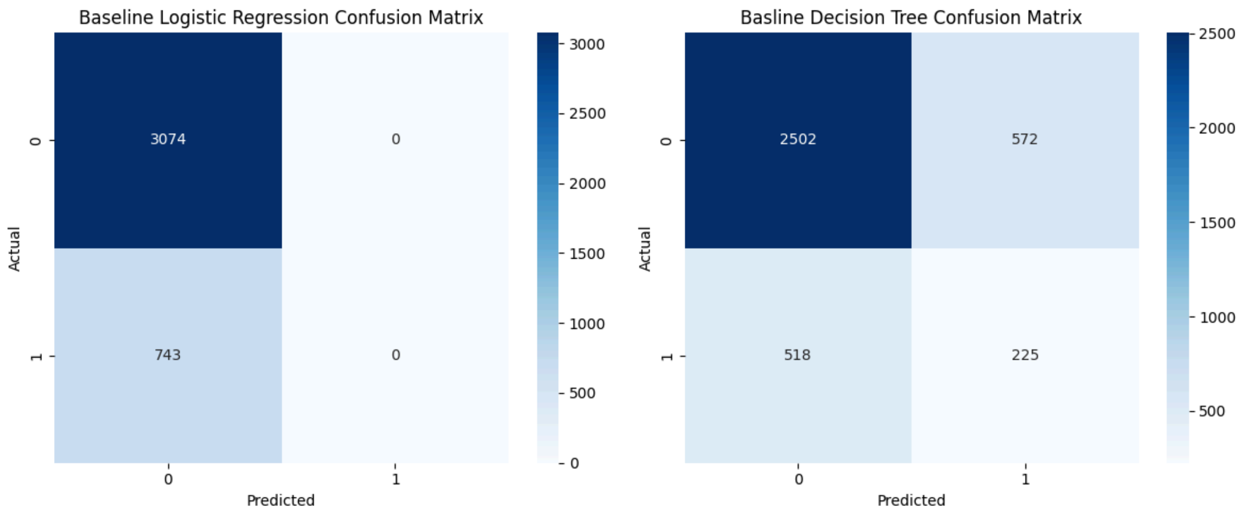


Figure 8. Confusion Matrix for Baselines of each model

Figure 8 shows the confusion matrix of baselines for each model. Logistic Regression predicts no suspicious users and the Decision Tree is able to correctly identify a couple of suspicious users. According to figure 9, if we decided to use accuracy we would find that logistic regression performed better but since it did not identify suspicious users we will use precision, recall, and F1 scores for class 1 which is a suspicious user. The metrics are defined as:

- *Precision:* The proportion of correctly identified suspicious users out of all users predicted as suspicious.

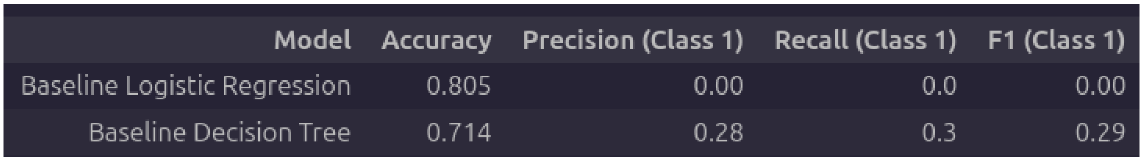
$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

- *Recall:* The proportion of correctly identified suspicious users out of all actual suspicious users.

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Negatives)}$$

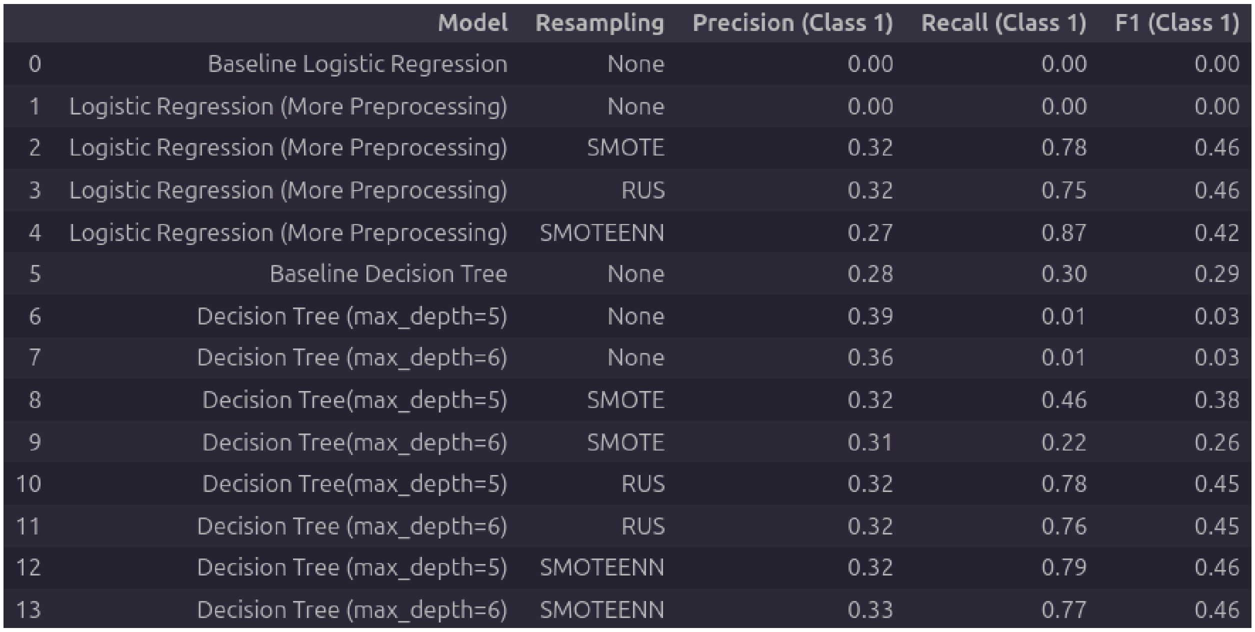
- *F1 Scale:* Provides a balanced measure of both Precision and recall.

$$F1\ Score = \frac{2(Precision * Recall)}{(Precision + Recall)}$$



Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1 (Class 1)
Baseline Logistic Regression	0.805	0.00	0.0	0.00
Baseline Decision Tree	0.714	0.28	0.3	0.29

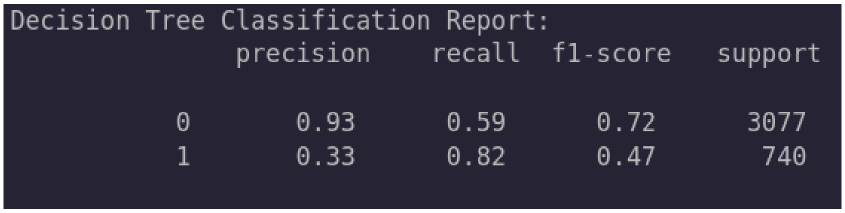
Figure 9. Metrics for Baselines of each model



	Model	Resampling	Precision (Class 1)	Recall (Class 1)	F1 (Class 1)
0	Baseline Logistic Regression	None	0.00	0.00	0.00
1	Logistic Regression (More Preprocessing)	None	0.00	0.00	0.00
2	Logistic Regression (More Preprocessing)	SMOTE	0.32	0.78	0.46
3	Logistic Regression (More Preprocessing)	RUS	0.32	0.75	0.46
4	Logistic Regression (More Preprocessing)	SMOTEENN	0.27	0.87	0.42
5	Baseline Decision Tree	None	0.28	0.30	0.29
6	Decision Tree (max_depth=5)	None	0.39	0.01	0.03
7	Decision Tree (max_depth=6)	None	0.36	0.01	0.03
8	Decision Tree(max_depth=5)	SMOTE	0.32	0.46	0.38
9	Decision Tree(max_depth=6)	SMOTE	0.31	0.22	0.26
10	Decision Tree(max_depth=5)	RUS	0.32	0.78	0.45
11	Decision Tree (max_depth=6)	RUS	0.32	0.76	0.45
12	Decision Tree (max_depth=5)	SMOTEENN	0.32	0.79	0.46
13	Decision Tree (max_depth=6)	SMOTEENN	0.33	0.77	0.46

Figure 10. Metrics after improving models

After making adjustments to data and model, figure 10 shows that with more preprocessing and SMOTE, Logistic Regression improved the baseline most with a F1 score of 0.46. A similar F1 score was achieved by the decision tree with max depth of 5 using SMOTEENN. There could be further adjustments made to improve both models. Since the decision tree had a very slight higher recall score, we will make one final adjustment.



```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.59      0.72      3077
           1       0.33      0.82      0.47       740
```

Figure 11. Decision Tree with a max depth of 3 using SMOTEENN

With a final adjustment of max depth to 3 and using SMOTEENN, we were able to achieve a F1 score of 0.47. Since this was the model that achieved the highest F1 score, we will create the tree to view its decisions.

Discussion

Figure 12 visualizes the decision-making process of the trained decision tree. The root node splits based on video view count: videos with 11,388 views or fewer are further evaluated based on the presence of the word 'friend' or video duration. If 'friend' is present (less than 1 occurrence) or the video is shorter than 27 seconds, the user is more likely classified as 'Not

Suspicious.' Videos with more than 11,388 views are subsequently assessed based on the occurrence of 'media' and 'internet'; the presence of either increases the likelihood of a 'Suspicious' classification. Since leaves contain features that were not part of feature importance from figure 7, a case can be made to remove more features from the dataset.

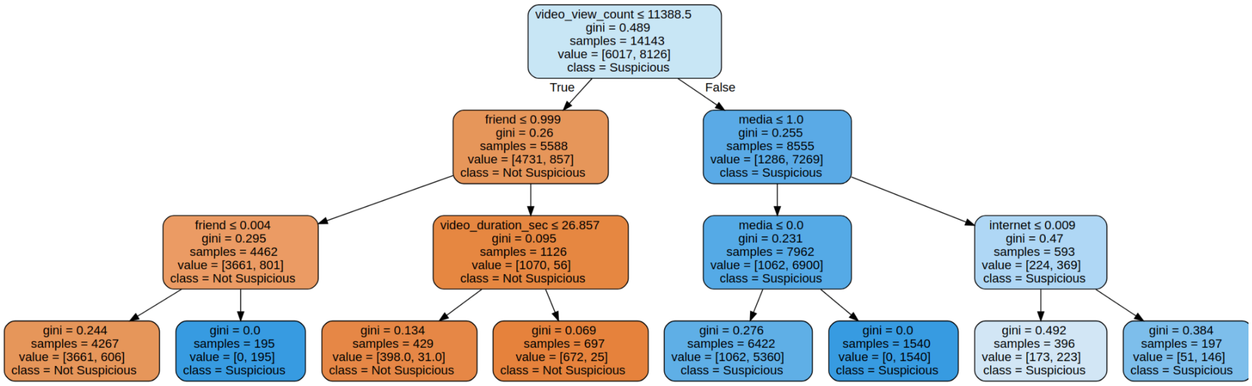


Figure 12. Decision Tree created by best performing model

With a F1 score of 0.47, it would not be acceptable to be deployed and used. Further refinement of the models may be able to achieve a score of 0.5 such as different resampling techniques, more advanced text processing such as TF-IDF or using word embeddings, or acquiring more data.

Limitations

The dataset exhibited significant class imbalance, which, even after using techniques like SMOTE and SMOTEENN, may have introduced synthetic bias. This could affect the generalizability of the models in real-world scenarios.

The models were trained on data specific to TikTok users, which may limit their applicability to other platforms or contexts. Additionally, the reliance on features like word frequency could be less effective in non-English or multi-lingual environments. Other text processing techniques were excluded from this research.

Recommendations

While prioritizing interpretability, simpler models like Logistic Regression and Decision Trees may not capture complex, latent patterns in user behavior, which more advanced models could identify.

Exploring this dataset with more complex models then using explainable AI (XAI) techniques to provide explanations why a model made a decision. According to Molnar (2024), post hoc methods could be applied to the models afterwards to give the complex model interpretability.

Conclusion

This research aimed to identify suspicious users on TikTok by analyzing user behavior patterns and features of the content published by the user. By prioritizing model interpretability, Logistic Regression and Decision Tree models were deployed and optimized to identify suspicious users, evaluate the accuracy of this identification, and interpret the decision-making process of the best-performing model. The initial baseline models struggled to identify suspicious users, with the Logistic Regression failing to identify any and the Decision Tree only correctly classifying a small number. Through data preprocessing techniques, including outlier removal, feature selection, and addressing class imbalance using SMOTE and SMOTEENN, significant improvements were made. The final Decision Tree model, with a maximum depth of 3 and utilizing SMOTEENN, achieved the highest F1 score of 0.47. This model's decision-making process was visualized, revealing key features such as video view count, the presence of specific words like "friend," "media," and "internet," and video duration as important factors in classifying users as suspicious. Despite these improvements, limitations such as dataset imbalance and model generalizability remain. Future research should explore more complex models combined with explainable AI techniques to potentially capture more intricate patterns while maintaining some level of interpretability.

Reference list

Afroogh, S., Akbari, A., Malone, E., Kargar, M. and Alambeigi, H. (2024). Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, [online] 11(1). doi:https://doi.org/10.1057/s41599-024-04044-8.

Dixon, S.J. (2024a). *Most Popular Social Networks Worldwide as of January 2024, Ranked by Number of Monthly Active Users*. [online] Statista. Available at: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Dixon, S.J. (2024b). *Number of Social Media Users Worldwide from 2017 to 2027*. [online] Statista. Available at: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

Gongane, V.U., Munot, M.V. and Anuse, A.D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1). doi:https://doi.org/10.1007/s13278-022-00951-3.

imbalanced-learn (2016). *imbalanced-learn documentation — Version 0.9.1*. [online] imbalanced-learn.org. Available at: https://imbalanced-learn.org/stable/index.html.

Molnar, C. (2024). *Chapter 9 Local Model-Agnostic Methods | Interpretable Machine Learning*. [online] Github.io. Available at: https://christophm.github.io/interpretable-ml-book/local-methods.html.

Shi, Z., Yao, W., Li, Z., Zeng, L., Zhao, Y., Zhang, R., Tang, Y. and Wen, J. (2020). Artificial intelligence techniques for stability analysis and control in smart grids: Methodologies, applications, challenges and future directions. *Applied Energy*, 278, p.115733. doi:https://doi.org/10.1016/j.apenergy.2020.115733.

Shokirov, Y. (2023). *TikTok User Engagement Data*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/yakhyojon/tiktok.

Tiktok. (2024). *Enforcement*. [online] Available at:
https://www.tiktok.com/community-guidelines/en/enforcement [Accessed 21 Dec. 2024].