TikTok Claims Classification Project

EXECUTIVE SUMMARY REPORT

Preliminary Data Summary

ISSUE/ PROBLEM

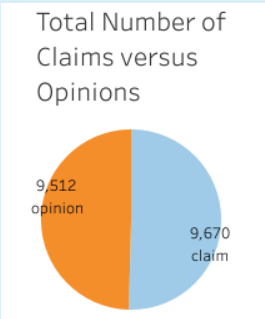
The TikTok data team aims to create a machine learning model for classifying claims within user submissions. To kickstart the process, the team must organize the raw dataset and preprocess it in preparation for future exploratory data analysis.

IMPACT

The findings from this preliminary analysis will significantly influence the subsequent steps. To comprehend the influence of user videos, the data team pinpointed two crucial variables for consideration: video_duration (in seconds) and video_view_count. These variables are deemed essential factors for future prediction models.

KEY INSIGHTS

- With the realization of a near equal balance between opinions and claims, we can confidently move forward with our future analysis, assured that the dataset contains a fairly even distribution of both types of content.
- Having identified the key variables and completed the initial investigation of the claims classification dataset, we are now prepared to embark on the exploratory data analysis phase.



RESPONSE

The data team conducted an initial exploration of the claims classification dataset, focusing on uncovering significant relationships between variables. In line with the objective of classifying user claims, they examined the distribution of claims and opinions to grasp the prevalence of each type of video content.

DETAILS

Upon examining the dataset, the claim_status variable emerged as notably pertinent to the client's proposed project. The subsequent screenshots highlight crucial analytical points essential for comprehending the claim_status variable.

```
data['claim_status'].value_counts()
```

```
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

- The data team delved into viewer engagement concerning videos categorized as claims and opinions. To gauge viewer engagement effectively, they scrutinized the view count. Analyzing both the mean and median view counts provided insights into the influence of each video category. Specifically, these statistics shed light on the correlation between content type (claim or opinion) and video views.

Claims:

Mean view count claims: 501029.45274771
Median view count claims: 501555.0

Opinions:

Mean view count opinions: 4956.43224989
Median view count opinions: 4953.0

Exploratory Data Analysis

OVERVIEW

The TikTok data team is embarking on a project to construct a machine learning model aimed at classifying claims within user submissions. This phase of the project necessitates thorough analysis, exploration, cleaning, and structuring of the data before proceeding with any model building efforts.

PROJECT STATUS

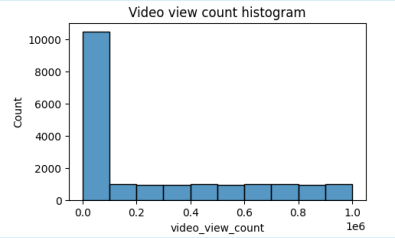
At this stage, the TikTok data team performed exploratory data analysis to delve into the impact of videos on TikTok users. Their objective was to comprehend user engagement by examining variables such as view count, like count, and comment count. These analyses aimed to shed light on the extent of interaction and interest generated by TikTok videos among users.

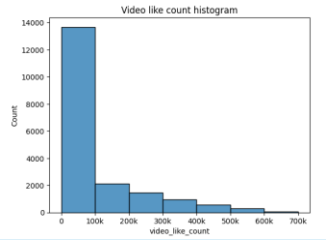
IMPACT

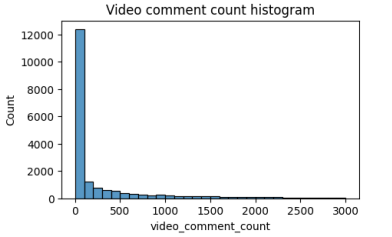
Based on the insights gleaned from the exploratory data analysis, it's evident that the future claim classification model must address null values and account for the imbalance in opinion video counts. To achieve this, these factors need to be integrated into the model parameters, ensuring robustness and accuracy in classification.

DETAILS

A vital aspect of the exploratory data analysis for this project entails visualizing the dataset. As depicted in the subsequent histograms, it's evident that a significant portion of videos are clustered towards the lower end of the value range for three variables representing user engagement with TikTok videos in this dataset.







KEY INSIGHTS

The exploratory data analysis carried out by TikTok's data team uncovered several crucial considerations for the classification model. These include handling missing values, balancing between "claims" and "opinions," and understanding the overall distribution of data variables. From this analysis, two key insights emerged:

- Null Values:

The exploratory data analysis identified over 200 null values across 7 distinct columns. Consequently, it's imperative for future modeling efforts to account for these null values to prevent drawing insights that rely on complete data. Further analysis is warranted to probe the underlying reasons for these null values and assess their potential impact on subsequent statistical analyses or model building endeavors.

- Skewed Data Distribution:

The exploratory analysis revealed a right-skewed distribution in video view and like counts for opinions, with concentrations predominantly in the lower range, around 1,000. This observation underscores the need to tailor the choice of models and model types to accommodate this skewed data distribution effectively.

Statistical Testing Results

OVERVIEW

The TikTok data team is aiming to create a machine learning model for classifying claims within user submissions. In this phase of the project, the team will undertake a hypothesis test to scrutinize the connection between verified_status and video_view_count.

FINDINGS

The analysis indicates a disparity in the number of views between TikTok videos posted by verified accounts and those posted by unverified accounts. Consequently, these findings imply potential fundamental behavioral distinctions between these two account groups: verified and unverified. It would be worthwhile to delve into the underlying reasons for this behavioral discrepancy. For instance, exploring whether unverified accounts tend to share more engaging content and whether this content pertains to claims or opinions could provide valuable insights. Additionally, investigating the possibility of unverified accounts being linked to spam bots, which could artificially inflate view counts, merits consideration.

DETAILS

The TikTok data team investigated the relationship between verified_status and video_view_count through two distinct approaches. Firstly, they examined the mean values of video_view_count for each group of verified_status within the sample data, revealing that the majority of accounts were unverified, with 265,663 unverified accounts compared to 91,439 verified accounts.

```
verified_status
not verified    265663.785339
verified         91439.164167
Name: video_view_count, dtype: float64
```

Secondly, a two-sample hypothesis test was conducted, consistent with the initial findings from the mean values analysis. This statistical examination corroborated that any observed difference in the sample data is attributable to a genuine disparity in the corresponding population means.

NEXT STPES

The team recommends advancing to construct a regression model focusing on verified_status. Such a regression model can provide insights into user behavior within the verified user group. This contextual understanding can then be leveraged to interpret outcomes from a subsequent claim classification model, which will be developed thereafter.

Regression Analysis

OVERVIEW

The TikTok data team is pursuing the development of a machine learning model to aid in the classification of claims for user submissions. Previously, the team noted a notable trend: verified users are considerably more inclined to post opinions. Given the overarching objective of classifying claims and opinions, it is imperative to construct a model capable of predicting the behavior of the verified account type, which tends to share more opinions. As a result, in this phase of the project, the data team proceeded to build a logistic regression model aimed at predicting verified_status.

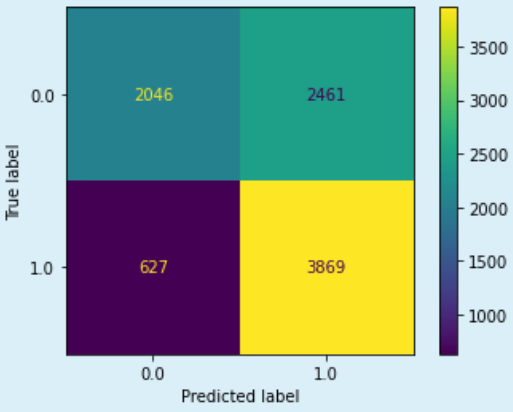
PROJECT STATUS

The choice of the verified_status variable for this regression model stemmed from the observed relationship between the verified account type and the nature of video content. Additionally, a logistic regression model was deemed suitable due to the data type and distribution.

A glance at the model results reveals that the logistic regression model attained a precision of 69% and a recall of 66% (weighted averages). Moreover, the model achieved an f1 accuracy of 66%. These findings offer significant insights into video features, which are further elaborated upon in the section titled "key insights."

DETAILS

According to the estimated model coefficients from the logistic regression analysis, longer videos are positively correlated with higher odds of the user being verified. However, other video features display small estimated coefficients in the model, suggesting a minor association with verified status. Consequently, besides video length, other video features do not appear to have a significant association with verified status.



NEXT STEPS

Moving forwards, the team need to build a classification model aimed at predicting the status of claims made by users. This represents the ultimate objective and initial expectation set forth by the TikTok team. With sufficient data and insights accumulated, the results of this model can now be analyzed with valuable context regarding user behavior.

Machine Learning Model

OVERVIEW

The TikTok data team is striving to create a machine learning model to aid in distinguishing videos as either claims or opinions. Prior examination of the available data unveiled a strong correlation between video engagement levels and claim status. With this insight in mind, the team is assured that the resultant model will fulfill all performance criteria.

PROJECT STATUS

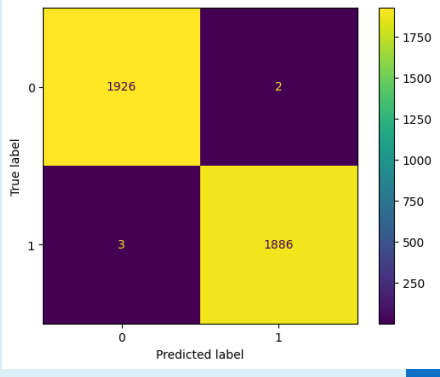
TikTok videos frequently garner numerous user reports for various reasons, yet not all reported videos can be individually reviewed by a human moderator. However, videos that assert claims, as opposed to merely expressing opinions, are significantly more likely to feature content that contravenes the platform's terms of service. In light of this, TikTok is seeking a method to identify videos that make claims in order to prioritize them for review.

SOLUTION

The data team constructed two tree-based classification models. Both models were deployed to predict on a held-out validation dataset, and the ultimate model selection was based on the one exhibiting the highest recall score. Subsequently, the chosen final model was applied to score a test dataset to estimate its performance for future use.

DETAILS

- Both model architectures, random forest (RF) and XGBoost, demonstrated exceptional performance. However, the RF model outperformed with a superior recall score of 0.995, earning it the title of champion model.
- Performance on the test holdout data yielded near perfect scores, with only five misclassified samples out of 3,817.
- Further analysis confirmed the anticipated trend: the main predictors were all linked to video engagement levels. Factors such as video view count, like count, share count, and download count were found to contribute substantially to the predictive signal within the data. Based on these findings, it can be deduced that videos with heightened user engagement levels were significantly more inclined to be claims. Interestingly, it was observed that no opinion video garnered more than 10,000 views.



NEXT STEPS

As previously mentioned, the model exhibited outstanding performance on the test holdout data. However, before deploying the model, the data team suggests conducting further evaluation using additional subsets of user data. Additionally, they recommend ongoing monitoring of the distributions of video engagement levels to ensure that the model remains resilient to fluctuations in its most predictive features. This proactive approach will help maintain the model's effectiveness and reliability over time.