Project background

*This is a full cycle machine learning project that is required to graduate the Google Advanced Data Analytics Certificate. Even though there is an exemplar for each stage of the project, I tried my first attempt to emerge myself to the real job scenario and compare with the exemplar code.

Welcome to your new role at TikTok! We're thrilled to have you on the data team!

At TikTok, our mission is to inspire creativity and bring joy. Our employees lead with curiosity and move at the speed of culture. Combined with our company's flat structure, you'll be given dynamic opportunities to make a real impact on a rapidly expanding company and grow your career.

TikTok users have the ability to report videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. This process generates a large number of user reports that are difficult to address quickly.

TikTok is working on the development of a predictive model that can determine whether a video contains a claim or offers an opinion. With a successful prediction model, TikTok can reduce the backlog of user reports and prioritize them more efficiently.

Team members
As a new data analyst, you'll work closely with a talented team of experienced data professionals. You'll also collaborate with TikTok co-workers outside the data team as your project requires.

Data team roles

- Willow Jaffey- Data Science Lead
- Rosie Mae Bradshaw- Data Science Manager
- Orion Rainier- Data Scientist

Cross-functional team members

- Mary Joanna Rodgers- Project Management Officer
- Margery Adebowale- Finance Lead, Americas
- Maika Abadi- Operations Lead

Meeting notes

Now that you are onboard, you have access to the company network and are set up with a company email account. You notice an email from your supervisor at the top of your inbox.

From: Rosie Mae Bradshaw

Subject: Your first task: Review meeting notes

If you are able to read this, then your company accounts have been created! It's the perfect time to get started with your first task. Last week, I attended an internal meeting with our leadership team about a new project we are about to begin. You'll receive more information in the next few days, but I would like you to be aware of some needs that were identified by our leadership. Here is an excerpt from the notes I took during the leadership team meeting. I've organized the points by the individual who made the comment:

Mary Joanna Rodgers- Project Management Officer

- The project will need a global document identifying deliverables & milestones.
- We will need to generate some visuals to share with TikTok executives.

Orion Rainier-Data Scientist

- We will need to have an understanding of the data we will be working with on this project.
- It will require someone on the team to inspect the dataset for missing data.
- The data team needs to determine what information will be the most useful in TikTok's data through exploratory data analysis (EDA).
- Eventually, our team will need to run some method of statistical testing.

Willow Jaffey- Data Science Lead

- The goal of the claim classification project is to build and deliver a reliable machine learning model.
- The team will need to determine the type of regression model for this project.
- Once we have a final model, I'll need to know the main talking points going into our presentation to the leadership team.
- The team will need to determine the best method of hypothesis testing for this dataset.

My thoughts and concerns…

- I think it's best to use Python for this project. I'll have someone on my team set that up as soon as we have the project proposal in place.
- I'd suggest the data team will need to check the assumptions made by any regression modeling developed for this project.
- The data team will identify and set up the project's workflow.

There will be more details sent to you very soon.

Welcome to the team,

Rosie Mae Bradshaw

Data Science Manager

TikTok

Task 1: Project proposal

TikTok's data team is in the earliest stages of the claims classification project. The following tasks are needed before the team can begin the data analysis process:

A project proposal identifying the following:

- Organize project tasks into milestones
- Classify tasks using the PACE workflow
- Identify relevant stakeholders

Email from Rosie Mae Bradshaw, TikTok's Data Analysis Manager

Subject: Welcome to the TikTok data team

From: Rosie Mae Bradshaw@TikTok

Welcome to the TikTok data team!

We are so glad to have you aboard.

I'd like to get you started on our latest project. I'll provide a little background that will give some context.

With the number of submissions and interactions on TikTok each day, it's challenging for our moderators to review each video, user comment, and content claim in a timely manner. You will assist us with the development of machine learning techniques that will assist in distinguishing between claims or opinions within content on the TikTok platform. The goal is to build a machine learning model with the capability to reliably classify user interaction data as a claim or opinion.

For your first task, please draft a plan of action for the team. Include questions we need to answer before we get started on the project, important details to consider at the beginning of the project, and the milestones we'll likely need throughout the duration of the project. To help get you started, I'll pass along some bullet points I've collected from stakeholders in an email.

I recall from your interview that you've completed a Google Careers Certificate on Advanced Data Analytics. I've heard great things about the Google certificates. I'm sure that you'll be able to use those skills to help get this latest project started.

If you need anything as you are settling in, feel free to reach out.

Rosie Mae Bradshaw

Data Analysis Manager

TikTok

Task 2: Prepare Data

Previously, you were asked to complete a project proposal by your supervisor, Rosie Mae Bradshaw. You have received notice that the project proposal submitted by the team has been approved and your team has been given access to TikTok's user data. To get clear insights, the data must be inspected, organized, and prepared for analysis.

TikTok's data team is in the earliest stages of the claims classification project. The following tasks are needed before the team can begin the data analysis process:

- Build a dataframe for the TikTok dataset
- Examine data type of each column
- Gather descriptive statistics

You discover two new emails in your inbox: one from your supervisor, Rosie Mae Bradshaw, and one from Willow Jaffey, the data team's Data Science Lead.

Email from Rosie Mae Bradshaw, Data Science Manager

Subject: help with coding notebook?

From: "Bradshaw, Rosie Mae" —rosiemaebradshaw@tiktok

Cc: "Rainier, Orion"—orionrainier@tiktok

Good morning,

I have a couple of updates on our latest project. The leadership team has approved the project proposal that we completed previously. Thanks for all of your great work so far. Additionally, I just received an email from our Project Management Officer, Mary Joanna Rodgers that the data team is clear to proceed.

Before we begin the process of Exploratory Data Analysis (EDA), we could really use your help with coding and prepping the data. During your interview you mentioned that you worked with Python specifically in the Google certificate program you completed. That experience sounds applicable here.

Orion Rainier (Cc'd above) started a Jupyter notebook with the relevant dataset (attached). Orion is very involved in the final stages of another project. I'm sure your assistance in completing the coding and setting up the notebook for the project would be greatly appreciated.

Orion, do you mind sharing the details?

Humblest regards,

Rosie Mae Bradshaw

Email from Orion Rainier, Data Scientist

Subject: RE: Help with coding notebook?

From: “Rainier, Orion”—orionrainier@tiktok

Cc: “Bradshaw, Rosie Mae” —rosiemaebradshaw@tiktok

Nice to meet you (virtually)!

Hope you have enjoyed your first few weeks!

With the project proposal approved, we are ready to begin the process of preparing the claim classification data. The goal of this project is to ultimately build a machine learning model that can streamline the claims process by identifying whether statements made in videos are claims or opinions.

A claim refers to information that is either unsourced or from an unverified source. For example, “The news reported that someone revealed that around 50% of the mined gold on Earth comes from one source.”

Opinions refer to the personal beliefs or thoughts of a group or an individual. Here’s an example, “In my opinion the most productive work day of the week is Tuesday.”

There are a number of data team members committed to adjusting the machine learning developed for the last project, so your help is greatly appreciated!

Until we finish the prior project, there is no need to do a full EDA on this data. We will get to that soon. Do you mind importing the data (attached) and reviewing it for the team? It would be fantastic if you could include a summary of the column Data types, data value nonnull counts, relevant and irrelevant columns, along with anything else code related you think is worth sharing/showing in the notebook? You’ll need to select a couple of variables to focus on. Include their minimum and maximum values. I haven’t looked closely at the data yet, but it would be really helpful if you can create meaningful variables by combining or modifying the structures given.

Thanks,

Orion Rainier

Data Scientist

TikTok

The Dataset provided is TikTok dataset.csv, this is a simulation dataset created in partnership with TikTok.

Data Dictionary

Column name	Type	Description
#	int	TikTok assigned number for video with claim/opinion.
claim_status	obj	Whether the published video has been identified as an “opinion” or a “claim.” In this dataset, an “opinion” refers to an individual’s or group’s personal belief or thought. A “claim” refers to information that is either unsourced or from an unverified source.
video_id	int	Random identifying number assigned to video upon publication on TikTok.
video_duration_sec	int	How long the published video is measured in seconds.
video_transcription_text	obj	Transcribed text of the words spoken in the published video.
verified_status	obj	Indicates the status of the TikTok user who published the video in terms of their verification, either “verified” or “not verified.”
author_ban_status	obj	Indicates the status of the TikTok user who published the video in terms of their permissions: “active,” “under scrutiny,” or “banned.”
video_view_count	float	The total number of times the published video has been viewed.
video_like_count	float	The total number of times the published video has been liked by other users.
video_share_count	float	The total number of times the published video has been shared by other users.
video_download_count	float	The total number of times the published video has been downloaded by other users.
video_comment_count	float	The total number of comments on the published video.

Task 3: Exploratory Data analysis (EDA)

Orion is pleased with the work you have already completed and is requesting your assistance with some Exploratory Data Analysis (EDA) and data visualization.

Email from Orion Rainier, Data Scientist

Subject: Tik Tok Claims Classification EDA & Vizzes

From: “Rainier, Orion”—orionrainier@tiktok

Cc: “Bradshaw, Rosie Mae” —rosiemaebradshaw@tiktok; “Jaffey, Willow” —willowjaffey@tiktok

Hi there,

Thanks for the amazing work you’ve done so far.

We’re ready to perform EDA on the data. Has Rosie Mae told you what the management team expects when it comes to EDA? If not, think of it as a “show your work” kind of report. They will want to see a Python notebook showing the structuring and cleaning you did, as well as any matplotlib/seaborn visualizations you plotted to help us understand the data. I would suggest at the very least a graph comparing claim counts to opinion counts, and a breakdown of “author ban status” counts. But whatever you think makes most sense works for us.

Additionally, the management team has recently asked all EDA to include Tableau visualizations. We’ve found these to be particularly helpful in status reports to the client and board members. For this data, I suggest a Tableau dashboard showing a simple claims versus opinions count, as well as stacked bar charts of claims versus opinions for variables like video view counts, video like counts, video share counts, and video download counts. Make sure it is easy to understand to someone who isn’t data savvy, and remember that the assistant director is a person with visual impairments. I understand you have some Tableau experience? Let me know if you need help with this.

By the way, I CC’d our Data Science Lead, Willow Jaffey, who is on the senior management team and will be reviewing and approving our analysis before the project manager reports it back to the client. @Willow, I just want to keep you informed on the progress!

Thanks!

Orion Rainier

Data Scientist

TikTok

Email from Willow Jaffey, Data Science Lead

Subject: RE: Tik Tok Claims Classification EDA & Vizzes

From: "Jaffey, Willow" —willowjaffey@tiktok

Cc: "Bradshaw, Rosie Mae" —rosiemaebradshaw@tiktok; "Rainier, Orion"—orionrainier@tiktok

Thank you, Orion!

Welcome to the team, so glad to have you.

Along with the Tableau dashboard and notebook, it would be really helpful if you included an executive summary of your analysis attached via email.

Appreciate your help!

Willow Jaffey

Data Science Lead

TikTok

Task 4: Statistical models

The TikTok data analytics team has completed the first three milestones of the claims classification project and is nearing the halfway point. So far, the team has completed a project proposal, and used Python to perform exploratory data analysis on the dataset for the claims classification project. The team also produced data visualizations in both Python and Tableau to share with stakeholders. The next step is to use statistical methods to analyze and interpret the claims classification data.

Email from Mary Joanna Rodgers, Project Management Officer

Subject: New Request - Hypothesis Test: Verified/Unverified Accounts

From: “Rodgers, Mary Joanna” —maryjoannarodgers@tiktok

Cc: “Rainier, Orion”—orionrainier@tiktok; “Jaffey, Willow” —willowjaffey@tiktok; “Bradshaw, Rosie Mae” —rosiemaebradshaw@tiktok

Hello Data Team!

Really excellent work so far. The leadership team is impressed with the results–especially the progress and insights shared on the last executive summary report! Thanks so much for the hard work.

On that note, they have requested an additional item to be added to the initial project scope. We are interested in whether there is a statistical difference in the data between verified and unverified accounts. Do you have any indication which variable would be most insightful to test in terms of verified and unverified accounts? And please include an executive summary for the results.

Many thanks!

Mary Joanna Rodgers

Project Management Officer

TikTok

Network with TikTok employees from a variety of teams and locations. Participate in TikTok Tuesdays, every Tuesday @2pm EST.

Task 5: Regression model

The TikTok team has reviewed the results of the hypothesis testing. TikTok's Operations Lead, Maika Abadi, is interested in how different variables are associated with whether a user is verified. Earlier, the data team observed that if a user is verified, they are much more likely to post opinions. Now, the data team has decided to explore how to predict verified status to help them understand how video characteristics relate to verified users.

Email from Rosie Mae Bradshaw TikTok's Data Science Manager

Subject: RE: Details on Regression Phase

From: "Bradshaw, Rosie Mae" —rosiemaebradshaw@TikTok

Cc: "Jaffey, Willow" —willowjaffey@tiktok; "Rodgers, Mary Joanna" —maryjoannarodgers@tiktok; "Rainier, Orion"—orionrainier@tiktok; "Abadi, Maika,"— maikaabadi@tiktok

We've decided to look into how to predict 'verified_status', which we believe will help us understand how video characteristics relate to verified users. To achieve this, the data team will build a logistic regression model using 'verified_status' as the outcome variable. The results of this milestone will inform us as we approach constructing the final claims prediction model.

Feel free to reach out with additional questions.

Many thanks,

Rosie Mae Bradshaw

Data Analysis Manager

TikTok

Task 6: Construct the machine learning model

The data team at TikTok are close to completing the claims classification project. Earlier, the data team completed a project proposal, and used Python to explore and analyze the claims dataset, created data visualizations, and conducted a statistical test. Most recently, the team built a logistic regression model to predict whether statements are claims or opinions within submissions, based on the 'verified_status' variable within the claims dataset.

Email from the Data Science Lead

Subject: Claim model considerations

From: "Jaffey, Willow" —willowjaffey@tiktok

Cc:; "Bradshaw, Rosie Mae" —rosiemaebradshaw@tiktok

Happy Monday!

You have done great work so far. We are in the final stage of the claims classification project and I am excited to see the final claims model and how it can help TikTok streamline the claims process overall.

I've determined that an effective final model could be approached in several ways, such as a Naive Bayes or tree-based model. However, having looked at the data analysis so far, I believe our most effective method for the sample dataset is a random forest model.

Once complete, please send an executive summary of your random forest model and results to Rosie Mae and myself so we can prepare to present the results to the leadership team. Be sure to include what was requested in the email from Mary Joanna, a summary of the variables used to make the prediction, and an idea of the accuracy of the model.

I look forward to seeing what you build!

Willow Jaffey

Data Science Lead

TikTok