Course 5 workplace scenarios

TikTok

Project goal:

The TikTok data team is developing a machine learning model for classifying claims made in videos submitted to the platform.

Background:

TikTok is the leading destination for short-form mobile video. The platform is built to help imaginations thrive. TikTok's mission is to create a place for inclusive, joyful, and authentic content–where people can safely discover, create, and connect.

Scenario:

The data team at TikTok is close to their goal of building a model to assist in the classification of claims in videos. The next step is to use the project data to create a regression model. As a member of TikTok's data team, you'll determine the type of regression model that is needed and develop one using TikTok's claim classification data.

Course 5 tasks:

- Import relevant packages and TikTok data
- Exploratory data analysis and check model assumptions
- Determine the correct modeling approach
- Build the regression model
- Finish checking model assumptions
- Evaluate the model
- Interpret model results and summarize findings for cross-departmental stakeholders within TikTok

Note: *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

Key Takeaways

In Course 5, Regression Analysis: Simplify Complex Data Relationships, you practiced modeling variable relationships, and investigated linear and logistic regression to better understand data modeling. Additionally, you reviewed model assumptions and evaluation techniques that will help you interpret and articulate relationships in datasets.

Course 5 skills:

- Conduct statistical analysis
- Conduct regression modeling
- Create predictive models
- Expand Python coding
- Share Insights and Ideas with stakeholders

Course 5 end-of-course project:

- Regression model within a Python notebook
- Executive summary with results of model and insights

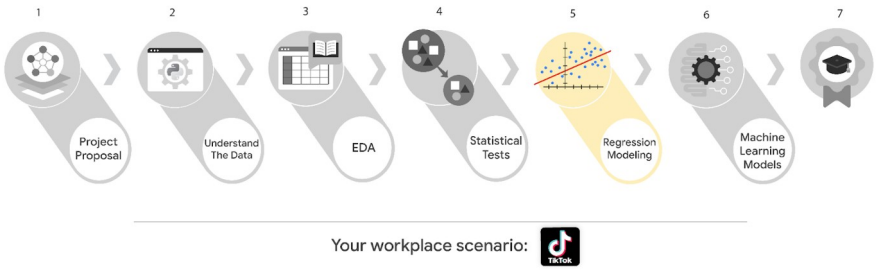
The end-of-course portfolio projects are designed for you to apply your data analytical skills within a workplace scenario. No matter which scenario you work with, you will practice your ability to discuss data analytic topics with coworkers, internal team members, and external clients.

As a reminder, you are required to complete one project for each course. To gain additional practice, or to add more samples to your portfolio, you may complete as many of the scenarios as you wish.

Course 5 end-of-course portfolio project overview: TikTok

Learn about the Course 5 TikTok workplace scenario!

The end-of-course project in Course 3 focuses on your ability to use exploratory data analysis to organize and understand the data within a project. As a reminder, in Course 1 you developed a project proposal that outlined milestones, which progress with each of the end-of-course projects. A visual representation is provided in the graphic shown here:



Learn more about the project, your role, and expectations in this reading.

Background on the TikTok scenario

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 submit reports that identify videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. The process generates a large number of user reports that are challenging to consider in a timely manner.

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.

Project background

TikTok's data team is working on the claims classification project. The following tasks are needed at this stage of the project:

- Determine the correct modeling approach
- Build a regression model
- Finish checking model assumptions

- Evaluate the model
- Interpret model results and summarize findings for cross-departmental stakeholders within TikTok

Your assignment

You will create a regression model for the claims classification data. You'll determine the type of regression model that is needed and develop one using TikTok's claim classification data.

Team members at TikTok

Data team roles

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

The members of the data team at TikTok are well versed in data analysis and data science. Messages to these more technical coworkers should be concise and specific.

Cross-functional team members

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

Your TikTok team includes several managers, who oversee operations. It is important to adjust your general correspondence appropriately to their roles, given that their responsibilities are less technical in nature.

Note: *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

Specific project deliverables

With this end-of-course project, you will gain valuable practice and apply your new skills as you complete the following:

- Course 5 PACE Strategy Document to consider questions, details, and action items for each stage of the project scenario
- Answer the questions in the Jupyter notebook project file
- Create a regression model
- Evaluate the model
- Create an executive summary to share your results

Key takeaways

The Google Advanced Data Analytics Certificate end-of-course project is designed for you to practice and apply course skills in a fictional workplace scenario. By completing each course's end-of-course project, you will have work examples that will enhance your portfolio and showcase your skills for future employers.

Activity Overview

In this activity, you will showcase your ability to use Python to build a multiple linear regression (MLR) model. You will also update team members and stakeholders through an executive summary, demonstrating your ability to organize and communicate key information.

For additional information on how to complete this activity, review the previous readings: *End-of-course project introduction* and *Course 5 end-of-course portfolio project overview: TikTok*.

Be sure to complete this activity before moving on. The next course item will provide you with completed exemplars to compare to your own work. You will not be able to access the exemplars until you have completed this activity.

Scenario

You are a member of the TikTok data analytics team. The team is currently more than halfway through the claims classification project. Earlier, you completed a project proposal, used Python to explore and analyze the claims classification dataset, created data visualizations, and conducted a hypothesis test.

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. Therefore, you have been asked to conduct a logistic regression using verified status as the outcome variable. The results may be used to inform the final model related to predicting whether a video is a claim vs an opinion.

You check your inbox and discover a new email from Maika Abadi asking the data team about the details of a regression model. You also notice two follow-up emails from Rosie Mae. The first lists the specific variables that should be analyzed in the logistic regression model. The second email asks you to help build the model and prepare an executive summary to share your results.

Note: *Team member names used in this workplace scenario are fictional and are not representative of TikTok.*

Email from Maika Abadi, Operations Lead

Subject: Details on Regression Model

From: "Abadi, Maika,"— maikaabadi@tiktok

Cc: ““Jaffey, Willow” —willowjaffey@tiktok; “Rodgers, Mary Joanna” —maryjoannarodgers@tiktok; “Bradshaw, Rosie Mae” rosiemaebradshaw@TikTok; “Rainier, Orion”—orionrainier@tiktok

Hello Data Team,

I really appreciate your work, and thanks for the explanation of the next phase of the data analysis.

I’m curious to know more about how different variables are associated with whether a user is verified. I was hoping to get a bit more detail on the regression your team is planning to conduct. Will you be applying a linear regression model or a logistic regression model? It wasn’t clear in the meeting, and I wanted to be sure our teams are aligned on expectations.

Thank you,

Maika Abadi

Operations Lead

TikTok

Learn about using TikTok to do good

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

Thank you for your email.

Apologies that it was not clear in the meeting.

To answer your question, 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

Learn about TikTok’s Trust & Safety team

Email from Rosie Mae Bradshaw TikTok’s Data Science Manager

Subject: RE: Details on Regression Phase

From: “Bradshaw, Rosie Mae” —rosiemaebradshaw@TikTok

Cc: “Rainier, Orion”—orionrainier@tiktok

Hello my Data team!

Would you two mind completing the following using the fictional test data set::

- Logistic regression model in a Python notebook based ‘verified_status’ variable in the claims classification dataset
    - Be sure to include a confusion matrix of the results and the accuracy score of the model
- Draft an executive summary of your results

I’d appreciate a chance to look it over before you send it over to Mary Joanna. Please write the summary as if you’re addressing the leadership team.

Best regards,

Rosie Mae Bradshaw

Data Analysis Manager

TikTok

Learn about TikTok’s Trust & Safety team

Step-By-Step Instructions

Follow the instructions to complete the activity. Then, go to the next course item to compare your work to a completed exemplar.

Step 1: Access the templates

To use the templates for this course item, click each link below and select *Use Template*.

Link to templates:

- Course 5 PACE strategy document
- Course 5 Executive summary

OR

If you do not have a Google account, you can download the templates directly from the attachments below:

Step 2: Access the end-of-course project lab

***Note**: The following lab is also the next course item. Once you complete and submit your end-of-course project activity, return to the lab instructions' page and click **Next** to continue on to the exemplar reading.*

To access the end-of-course project lab, click the following link and select *Open Lab*.

- Course 5 TikTok project lab

Your Python notebook for this project includes a guided framework that will assist you with the required coding. Input the code and answer the questions in your Python notebook to build a regression model. You'll find helpful reminders for tasks like:

- Model building and evaluation
- Checking model assumptions
- Interpreting model results

You will also discover questions in this Python notebook designed to help you gather the relevant information you'll need to write an executive summary for your team.

Use your completed PACE strategy document and Python notebook to help you prepare your executive summary in the next step.

Data Dictionary

This project uses a dataset called tiktok_dataset.csv. It contains synthetic data created for this project in partnership with TikTok. Examine each data variable gathered.

19383 rows – Each row represents a different published TikTok video in which a claim/opinion has been made.

12 columns

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

Column name	Type	Description
		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

Step 3: Complete your PACE strategy document

The **Course 5 PACE strategy document** includes questions that will help guide you through the Course 5 TikTok project. Answer the questions in your PACE strategy document to prepare to use Python to inspect and organize your data.

As a reminder, the PACE strategy document is designed to help you complete the contents for each of the templates provided. You may navigate back and forth between the PACE strategy document and the Python notebook. Make sure your PACE strategy document is complete before preparing your executive summary.

Step 4: Prepare an executive summary

Your executive summary will keep your teammates at TikTok informed of your progress. The one-page format is designed to respect teammates and stakeholders who may not have time to read and understand an entire report.

First, select one of the executive summary design layouts from the provided template. Then, add the relevant information. Your executive summary should include the following:

- A summary of the variables analyzed in your regression model
- The results of your analysis
- Recommendations or insights based on your results

Complete your executive summary to effectively communicate your results to external stakeholders.

Pro Tip: Save the templates

Finally, be sure to save a blank copy of the templates you used to complete this activity. You can use them for further practice or in your professional projects. These templates will help you work through your thought processes and demonstrate your experience to potential employers.

What to Include in Your Response

Later, you will have the opportunity to self assess your performance using the criteria listed below. Be sure to address the following elements in your completed activity.

Course 5 PACE strategy document:

- Answer the questions in the PACE strategy document

Course 5 Python notebook:

- Build a regression model

Course 5 Executive summary:

- Include regression assumptions
- Identify the outcome and impact of your work for this data project