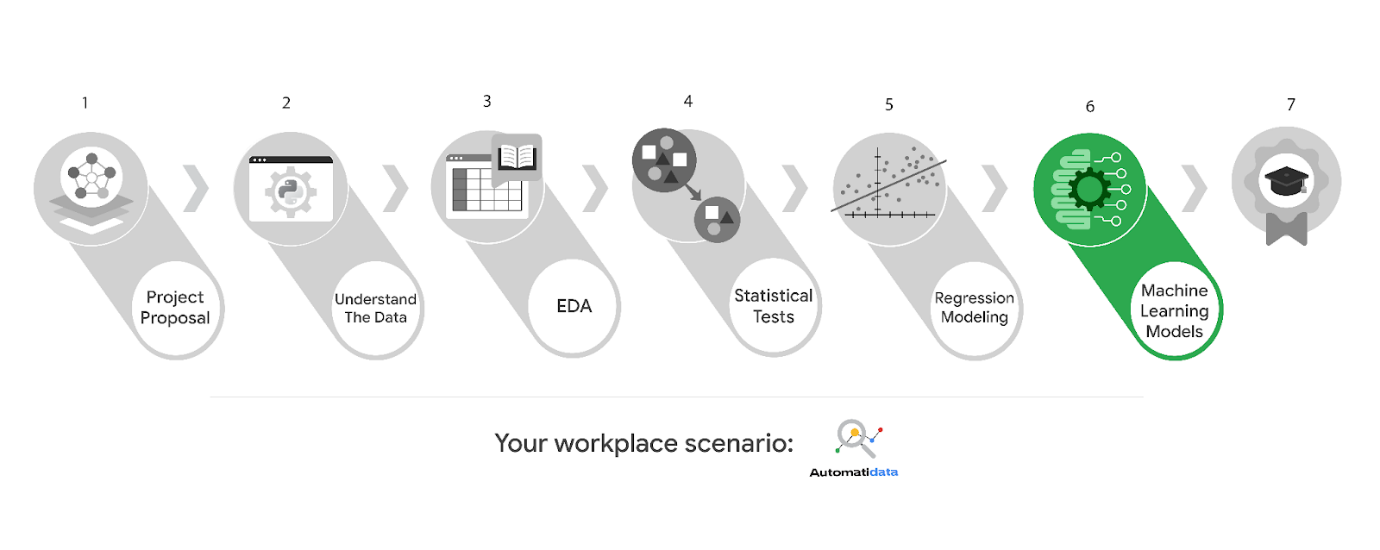


**Course 6 tasks:**

To complete this task, you’ll need to:

* Build a random forest model from the New York City TLC dataset
* Create an executive summary for the Automatidata data team before sharing the results with the client

The end-of-course project in Course 6 focuses on your ability to build a machine learning model for 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.

**Project background**

Automatidata is ready to create a machine learning model for TLC. The following tasks are needed to complete the project:

* Model building
* Model evaluation
* Summarize findings for Automatidata and the stakeholders at TLC

You will create a machine learning model for the TLC data. You will be responsible for leading these tasks, which include feature engineering, model development, and evaluation.

**Team members of Automatidata and the New York City TLC**

**Automatidata Team Members**

* Udo Bankole, Director of Data Analysis
* Deshawn Washington, Data Analysis Manager
* Luana Rodriquez, Senior Data Analyst
* Uli King, Senior Project Manager

Your teammates at Automatidata have technical experience with data analysis and data science. However, you should always be sure to keep summaries and messages to these team members concise and to the point.

**New York City TLC Team Members**

* Juliana Soto, Finance and Administration Department Head
* Titus Nelson, Operations Manager

The TLC team members are program managers who oversee operations at the organization. Their roles are not highly technical, so be sure to adjust your language and explanation accordingly.

**Specific project deliverables**

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

* Complete a PACE planner mapping questions, details, and action items for each stage of the project scenario
* Answer the questions in the Jupyter notebook project file
* Design and implement of a machine learning model
* Draft an executive summary of your results

**Scenario**

You are the newest member of Automatidata’s data analytics team. Your team is close to completing their project for the New York City Taxi & Limousine Commission (TLC). Previously, you completed a project proposal and used Python to explore and analyze the TLC dataset, create data visualizations, and conduct an A/B test. Most recently, you built an MLR model for fare amounts based on a variety of variables.

The New York City TLC is impressed with your work so far. Now, they want your team to identify which variables or factors influence the amount of gratuity a rider gives a driver. Your work will help TLC stakeholders make informed business decisions that will increase gratuities and subsequently improve driver satisfaction.

At a meeting with New York City TLC stakeholders, your team suggests building a random forest model to predict whether or not a rider will be a generous tipper (>= 20%). At the end of the meeting, Titus Nelson, the Operations Manager at the New York City TLC, says that he will share the suggestion with his organization's leadership team.

A few days after the meeting, you receive an email from Juliana Soto, a Department Head at the New York City TLC. Juliana says that TLC leadership likes the idea of using a random forest model to predict gratuity and asks the team to share more details about the model. You also receive a follow-up email from Udo Bankole, the Director of Data Analysis at Automatidata. Udo asks you to build the random forest model and to prepare an executive summary to share your results.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Email from Juliana Soto, Finance and Administration Department Head (NYC TLC)**

**Subject:** NYC TLC Approval of Algorithm

**From:** “Juliana Soto,”Juliana.Soto@tlc.nyc

**Cc:** “Udo Bankole,” Udo@automatidata; “Uli King” Uli@automatidata; “Deshawn Washington,” Deshawn@automatidata; “Luana Rodriguez” Luana@automatidata; “Titus Nelson,” Titus.Nelson@tlc.nyc

Hello Automatidata Team,

Thank you for providing the details for the final phase of the prediction algorithm we have requested. I apologize for missing many of the weekly project meetings, but Titus has kept me informed of your progress. We discussed in detail your proposal for using a random forest model for prediction, and we are in agreement with you.

If you would, please commence with the creation of the algorithm. It would be very helpful to provide a summary of what data indicators the algorithm is basing its results on and an idea of the confidence your team has in the accuracy of the result.

Thank you for your great work,

Juliana Soto

Finance and Administration Department Head

---

New York City Taxi & Limousine Commission

*Learn more about* [*TLC’s accessible vehicle initiatives*](https://www1.nyc.gov/site/tlc/about/accessibility.page)*.*

**Email from Udo Bankole, Director of Data Analysis (Automatidata)**

**Subject:** RE:NYC TLC Approval of Algorithm

**From:** “Udo Bankole,” Udo@automatidata

**Cc:** “Luana Rodriguez” Luana@automatidata;

Hello data pros!

You have done great work so far. We are excited to find out what else you can discover in the data and for you to help us make data-driven business decisions.If you would please build the random forest model we discussed using the data New York City TLC has provided. As you’re aware, you have already cleaned and run this data through a MLR model, but you always need to validate your variables and data. So please revisit the dataset.

Once complete, please send an executive summary to Deshawn and myself of what wording you plan to send to the client. Be sure to include what Juliana requested, a summary of the variables used, and an indication of how we can test the accuracy of the model.

I look forward to exploring your build!

Udo Bankole

Director of Data Analysis

Automatidata

**Step-By-Step Instructions**

**Step 1: Access the templates**

To use the templates for this course item, click the following links and select *Use Template*.

Link to templates:

* [Course 6 PACE strategy document](https://docs.google.com/document/d/1hPtIs4X7c5xmLSi8qs7Og2FEQHkELXBC_pGuJI1jF9o/template/preview?resourcekey=0-mSL0tC7opaF8XIOdXa1JIw)
* [Executive summary templates](https://docs.google.com/presentation/d/1Pps5GKxi1V31y2oRHRzU-xhJubkEYzCgEIfNjlEY3Og/template/preview)

Step 2: Access the end-of-course 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 random forest model. You’ll find helpful reminders for tasks like:

* Ethical considerations
* Feature engineering
* Model building and evaluation

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 2017\_Yellow\_Taxi\_Trip\_Data.csv. It contains data gathered by the New York City Taxi & Limousine Commission. For each trip, there are many different data variables gathered.

The dataset contains:

**408,294 rows** – each row represents a different trip

**18 columns**

| **Column name** | **Description** |
| --- | --- |
| ID | Trip identification number. |
| VendorID | A code indicating the TPEP provider that provided the record.  **1= Creative Mobile Technologies, LLC;**  **2= VeriFone Inc.** |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| tpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| Passenger\_count | The number of passengers in the vehicle.  This is a driver-entered value. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | TLC Taxi Zone in which the taximeter was engaged. |
| DOLocationID | TLC Taxi Zone in which the taximeter was disengaged. |
| RateCodeID | The final rate code in effect at the end of the trip.  **1= Standard rate**  **2=JFK**  **3=Newark**  **4=Nassau or Westchester**  **5=Negotiated fare**  **6=Group ride** |
| Store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before being sent to  the vendor, aka “store and forward,” because the vehicle did not have a connection to the server.  **Y= store and forward trip**  **N= not a store and forward trip** |
| Payment\_type | A numeric code signifying how the passenger paid for the trip.  **1= Credit card**  **2= Cash**  **3= No charge**  **4= Dispute**  **5= Unknown**  **6= Voided trip** |
| Fare\_amount | The time-and-distance fare calculated by the meter. |
| Extra | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| MTA\_tax | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| Improvement\_surcharge | $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015. |
| Tip\_amount | Tip amount – This field is automatically populated for credit card  tips. Cash tips are not included. |
| Tolls\_amount | Total amount of all tolls paid in trip. |
| Total\_amount | The total amount charged to passengers. Does not include cash tips. |

Step 3: Complete your PACE strategy document

The **Course 6 PACE strategy document** includes questions that will help guide you through the Course 6 Automatidata project. Answer the questions in your PACE strategy document to prepare for using Python to model 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 Automatidata teammates and New York City TLC stakeholders informed of your progress.The one-page format is designed to respect teammates and stakeholders who might 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 benefits and limitations of your random forest 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.

**What to Include in Your Response**

**Course 6 PACE strategy document**:

* Answer the questions in the PACE strategy document

**Course 6 Automatidata project lab**:

* Build a random forest model

**Course 6 executive summary**:

* Clearly articulate the challenges presented in this data project
* Identify the outcome of your work
* Include recommendations for future work/next steps