Please below the assignment,

**Pre-Course Model Development**

In this assessment, you are tasked with using the Sydney Transportation dataset to build a logistic regression model. Your goal in this assignment is to build the best model you can based on AUC (area under the curve). Use choice as your response variable (y-variable) and [roc\_auc\_scoreLinks to an external site.](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.metrics.roc_auc_score.html%23sklearn-metrics-roc-auc-score&data=05%7C01%7CKamila.Tocicka%40gds.ey.com%7Cb56599acc92645f7491d08db94d5b53b%7C5b973f9977df4bebb27daa0c70b8482c%7C0%7C0%7C638267418075506486%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=fgl%2FPUF5rzOPep1OR0kOIDNNzS%2Bxkzy4IOqoNLLo%2FsI%3D&reserved=0" \t "_blank) to calculate AUC. This assessment is pass/fail and is designed to gauge your readiness for model development in Python.

Dataset: [sydney\_transportation.xlsx](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fmycourses.hult.edu%2Fcourses%2F3433435%2Ffiles%2F237111731%3Fwrap%3D1&data=05%7C01%7CKamila.Tocicka%40gds.ey.com%7Cb56599acc92645f7491d08db94d5b53b%7C5b973f9977df4bebb27daa0c70b8482c%7C0%7C0%7C638267418075506486%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=5M3bh6tbtncomg%2Ba%2BPnOhTR7etaXChIt8Gf8TrJQje8%3D&reserved=0" \o "sydney_transportation.xlsx" \t "_blank)[Download sydney\_transportation.xlsx](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fmycourses.hult.edu%2Fcourses%2F3433435%2Ffiles%2F237111731%2Fdownload%3Fdownload_frd%3D1&data=05%7C01%7CKamila.Tocicka%40gds.ey.com%7Cb56599acc92645f7491d08db94d5b53b%7C5b973f9977df4bebb27daa0c70b8482c%7C0%7C0%7C638267418075506486%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=OrO6Gp5hpbm2QoV994GGZkp13kdP62hgVSqhV6S3xh8%3D&reserved=0)

Below is the encoding for choice:

* 0 = train
* 1 = car

1. **A) Deliverable:**

Jupyter Notebook or .zip file containing your Jupyter Notebook. Please use the following naming convention: **FamilyName\_FirstName\_PreAssessment**

1. **B) Criteria for Points**

This submission is worth 100 points (70 points needed to pass), based on the following tasks.

**Imports (10 points)** - Import the dataset and any libraries you will need for this project (pandas, sklearn, etc.).

**Exploratory Data Analysis (20 points)** – Explore the data and generate each of the following:

* Histograms for all columns in the dataset. \*
* Linear model plots for all x-features (from the seaborn library). Use choice as your value for the y \*
* A correlation matrix for all columns in the dataset. Comment on two of the x-features you believe will perform well in predicting choice. Make sure to include a nontechnical rationale for each x-feature. (5-10 sentences)

\*For these visuals, write about anything you find interesting. If you do not find anything interesting in a particular visual, please comment out its code. (3-5 sentences)

**Feature Engineering (20 points)** – Engineer at least two new features (in other words, create two new x-variables). Write 3-5 sentences on why you believe each newly-engineered feature will add value in predicting choice.

**Preprocessing (10 points)** – Transform any features that are heavily skewed. You may also (optionally) standardize the data using StandardScaler() or a similar method.

**Train-Test Split (10 points)** – Split the data into training and testing sets. Set your test size to 0.20 and stratify the response variable.

**Model Development (10 points)** – Develop a logistic regression model using LogisticRegression from scikit-learn.\*

* Instantiate a LogisticRegression model object.
* Fit to the training set.
* Separately score on both the training and testing sets using AUC as your scoring criterion.

\* Note that you are welcome to explore and develop other model types. However, you must generate a logistic regression model as stated above.

**Confusion Matrix (10 points)** – Generate a confusion matrix of your results.

**Comments (5 points)** – A minimum of one quality "#" comment for every 5 lines of code.

**Error Free (5 points)** – The submission is error free.

Note that DataCamp provides a great framework for model building in their course [Supervised Learning with scikit-learnLinks to an external site.](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fapp.datacamp.com%2Flearn%2Fcourses%2Fsupervised-learning-with-scikit-learn&data=05%7C01%7CKamila.Tocicka%40gds.ey.com%7Cb56599acc92645f7491d08db94d5b53b%7C5b973f9977df4bebb27daa0c70b8482c%7C0%7C0%7C638267418075506486%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Bb1TbAZ1qVrrEmrKbZ3KaBux6M4RWwwlhNFvuCeZaAY%3D&reserved=0). Please review this course as much as needed. You can access all courses on DataCamp free of charge and should have received an email from DataCamp with instructions.