Introduction

Welcome to the AT&T data science assessment. Do not collaborate with anybody on this assessment. The expectation is that the work you present is entirely your own.

This assessment is designed to let you show us three aspects of the kind of work a data scientist might perform at AT&T: 1) writing code to analyze data, 2) creating a model to provide useful insights and 3) presenting your findings to a non-technical audience.

You will submit the following three deliverables for this assessment. You should assume that your work could be passed on to a colleague who will maintain and update your code over time. Write and code accordingly.

1. **Code that shows your analysis of the data and answers the questions posed below.** Your code should be clean, well-written, readable, and easily maintainable by a future data scientist. Your choice of development environment is up to you, but you need to be prepared to share it in a teleconference app like MS Teams, Zoom or Webex. Your code should be mainly in Python; use of Spark/PySpark is fine.
2. **Your notes on your findings and answers to the questions in this assessment.** Assume your audience is technical and understands both statistics and data science techniques. Your notes do not need to be written or structured in a formal style, but they should be well organized, easy to present, and fully answer the questions in the assessment.
3. **A slide deck (PowerPoint or other) intended for a non-technical audience.** For this deliverable, imagine you will be presenting to a business audience who understands basic statistics, but prefers looking at bullet points and charts rather than code. Your presentation should not exceed 10 minutes. You may be asked to present this deck during the interview process. Be prepared to answer questions about your presentation.

When you are done, package up your code, your notes, and your slides into a .zip file and email it back to your recruiting manager.

Analysis Questions

You will be working with loan data published by Lending Club, an innovative, peer-to-peer lending company. The data can be downloaded here, and a description of the data will be provided in a separate attachment.

1. **Download and read the data** into your tool of choice. This should be done programmatically so future maintainers of the code don't need to download the data manually. The answer to this question should be in your code deliverable but does not need to be included in your notes or slides.
2. Find the **average annual income and number of loan applicants** by state. The answer to this question should be in your code and notes deliverables but does not need to be included in your slides.
3. Consider the annual income of applicants from West Virginia and New Mexico.
    a. Plot a **histogram** comparing the annual income of applicants from these states.
    b. Form and test a **hypothesis** regarding the average annual incomes of the residents of West Virginia and New Mexico.

    The answer to this question should be in your code deliverable but does not need to be included in your notes or slides.
4. We're interested in **predicting which applicants will default** on their loan.
    a. Build a **derived variable** representing whether an applicant defaulted on their loan. Consider a loan that is "Charged Off" as a default, and a loan that is "Fully Paid" to not have defaulted. The answer to this question should be in all three deliverables.
    b. Build a **binary classification model** to predict which loans will default. Your model should take a data set of applicants as input and return the probability of default for each applicant. You should thoroughly describe how you developed and validated your model and explain any assumptions you made. The answer to this question should be in all three deliverables.
    c. Imagine your client is considering entering the lending market but is very risk averse (they prefer low default rates even if it means accepting

lower rates of return). **Develop a strategy** for entering this market. Some things to consider:

   i. Which locations should we target?
   ii. To which segments of the population should we advertise?
   iii. Any other helpful strategies you can think of to keep default rates low?

The answer to this question should be in all three deliverables.

5. Show us what you can do! This is an **optional free-form** section. Do you have a data science skill that can separate you from other candidates? This is your chance to show it off. Here are some ideas to get you started, but you can answer this section any way you'd like.

   a. **Anomaly detection.** Can you spot any interesting anomalies or outliers in the data?

   b. **Visualization.** Create a visualization that reveals something interesting in the data or describes the data in a compelling way.

   c. **Kaggle Grand Master-level model performance.** Do you relish squeezing every last thousandth out of your F1 score? Develop a highly tuned, highly accurate model. Be sure to explain your approach to tuning and evaluating your model.

   d. **Data augmentation.** Can you find other public data sets that may improve your model? Ingest the data, join it to the original data set and explain how the new data did or did not improve your model's performance.

   e. **Be creative.** Come up with something we didn't think of and impress us with your amazing findings!