Dataiku Data Scientist Technical Assessment and Presentation

Background Information

The United States Census Bureau leads the country's Federal Statistical System; its primary responsibility is to collect demographic and economic data about America to help inform strategic initiatives. Every ten years, the census is conducted to collect and organize information regarding the US population with the intention of effectively allocating billions of dollars of funding to various endeavors (e.g., the building and maintenance of hospitals, schools, fire departments, transportation infrastructure, etc.). Additionally, the collection of census information helps to examine the demographic characteristics of subpopulations across the country.

The Data

You have been provided a sample dataset from the US Census archive containing detailed, but anonymized, information for ~300,000 individuals. This archive contains four files:

1. `census_income_learn.csv` (data for model training).
2. `census_income_test.csv` (data for model testing).
3. `census_income_metadata.txt` (metadata for both datasets).
4. `census_income_additional_info.pdf` (supplemental information).

You may download this data archive here.

Problem Statement

For this technical assessment, you have been tasked with identifying characteristics that are associated with a person making more or less than $50,000 per year; the target variable for your research question is the final column of the datasets.

As the data scientist on this project, you are to attempt to answer this question by constructing a data analysis/modeling pipeline. Code submissions should be in Python and making the solution easily readable and replicable by the team will give you additional marks. In the event you would like to use a different language or tool, please ask. Considerations for your data analysis should include, but are not limited to, the following:

- Exploratory Data Analysis: Numerical and/or graphical representations of the data that may help inform insights and/or tactics for answering the research question of interest.
- Data Preparation: Data cleaning, preprocessing, feature engineering, etc., that may aid in improving data clarity & model generation.
- Data Modeling: The building of a few competing models to predict the target variable.
- Model Assessment: A selection of the best model based on performance comparisons.

- Results: A concise summary of key findings, recommendations, & future improvements.

Presentation Guidelines

Data Scientists at Dataiku are customer facing. Our primary mission is to coach and support our customers in the use of Data Science Studio and help them become proficient users. This sometimes includes co-development on customer projects. Given this, your presentation is intended not only to gauge your general data science proficiency, but also considers your presentation and customer support skills. In this role **effective communication is critical.**

For this assessment you will prepare a solution and a presentation for the "customer". In the presentation, one or more Dataiku personnel will play the role of the customer.

You will be evaluated based on the following criteria:

- The technical quality of the solution presented
- Your conceptual understanding of common data science topics
- The effectiveness of the presentation
- The presentation skills demonstrated
- Your answers to the customer's questions

The presentation should be prepared to be roughly 20 minutes with an additional 20 minutes for questions and discussion of the solution. You may present your slides in any fashion you deem fit and should expect to explain your methods and results in a manner such that a non-technical audience would understand. You should assume that the customer audience may range from expert to novice in terms of data science, so the presentation should be geared towards a general audience with enough detail to satisfy experts and also general enough to clearly explain to the novice. You should be prepared to discuss any technical details, subjective choices, and assumptions you made in the assessment.

Remember, the goal of the exercise is not to necessarily solve the problem completely, but rather to illustrate a thought process, thoroughly explain an approach, and discuss and critique the methodology used to answer the research question of interest in a collaborative setting.

Submission Guidelines

You should submit all documents deemed relevant your personal GitHub repository:

1. Code files
2. Slides

Candidates will be assessed first based on their submission (presentation, code) and if they are successful we will schedule a role play interview with a technical team. As a result it is important that the notes in the code, the findings from the analysis and the slides are fairly self contained. However we appreciate that you will want to expand more on points during a full interview so bullet points are sufficient. More value is placed for the production readiness of the code.

If a candidate passes this stage they will be invited to a follow up interview where they will present their technical assessment findings to our data science team and more general technical questions.

Some Advice

Keep in mind that any data science project can continue for an eternity — there will always be more that could be explored. While you are not timed for this assessment, aim to spend a few hours constructing your submission with a particular focus on explaining the benefits and detractions to your approach. A word of advice:

“Do not let perfect be the enemy of the good.” -Voltaire

Please do not hesitate to reach out should you have any questions.

Lastly — good luck!