Interview Project for an Associate Data Scientist

As an Associated Data Scientist, you will be joining a team who is on a mission to disrupt the logistics industry by using statistical knowledge with machine learning plus other algorithmic techniques to drive business impact and take our technology to new heights with data science. For this interview project, you will be asked to play the role of an associate data scientist taking the hypothetical scenario below and transforming it into a model to automate the manual order acceptance process at C.H. Robinson.

Problem statement:

C.H. Robinson works with customers looking to ship goods with carriers, i.e. trucking companies or owner-operators of semi-trucks, to drive those goods from an origin location to a destination location. Customers pay C.H. Robinson money -- we call this a *rate*. C.H. Robinson in turn pays carriers to ship that freight -- we call this our *cost*. One outcome we try to avoid is having to pay a carrier more than the customer pays us, because our profit is roughly *rate – cost.*

When customers need us to ship freight, they request an *order*. C.H. Robinson, for a variety of reasons, may decide to reject the order, accept the order, or create another status for the order. An order contains the following freight characteristics, which are columns in your dataset: ??

Variable	Variable Type	Description
request_id	Categorical	The ID of the order request
week_id	Categorical	Integer for week of year
weekday	Categorical	String for day of week
miles	Numeric	The miles between the origin location and destination location
order_equipment_type	Categorical	What type of semi-truck is requested for the shipment
order_distance	Numeric	The miles the customer believes the truck will drive
order_num_stops	Categorical	The number of stops (pickups plus drop-offs) the carrier will have to make
order_origin_weight	Numeric	The weight of the shipment
lead_days	Numeric	The number of days before the pickup date that the customer placed the order
color	Categorical	A string used to visualize order status
origin_dat_ref	Categorical	An ID representing the geographic region of the origin location
dest_dat_ref	Categorical	An ID representing the geographic region of the destination location
rate_norm	Numeric	The *normalized* money the customer will pay us for this shipment
est_cost_norm	Numeric	The *normalized* ***estimated*** cost that C.H. Robinson will pay to a carrier
CurrentCondition	Categorical	The status of the order

Instructions:

1. Please use *IMA_recommendation_simulation_data.csv* that was provided.
2. Create a Jupyter notebook that follows the **Task** instructions below.
    a. The notebook can be in R or Python.
    b. Organize your notebook with headings using Markdown cells (i.e. Data and Notebook Setup, Exploratory Data Analysis, etc.)
    c. Your code must be included and kept in separate cells than your answers to the questions in each section.
    d. Visualizations can be produced directly from your code.
3. Upload your Jupyter notebook in a Github repository. **Note:** Please make sure the notebook code, visualization, and markdown cells display correctly in Github.
4. Submit your project using the web form here **within one week** of receiving the project.

Task:

Using the given data set, build **two** classifiers that predict the target variable, CurrentCondition, answering the following questions along the way.

Data and notebook setup

1. Load/read the data from the csv file into your Jupyter notebook.
    a. What is the shape of the data file you were given?
    b. How many unique categories are there for CurrentCondition?

Exploratory Data Analysis

2. Exploratory Data Analysis is a critical part of a Data Scientists job
    a. Is request_id unique?
    b. Is there any missing data in this data set? If so, which columns have missing data?
    c. Are there any features you would consider highly correlated? Please limit this to two pairs and describe why you selected the two pairs. Demonstrate this correlation numerically and visually.
    d. Show two other visualizations you would use during the EDA portion of the analysis. Describe your interpretation of the two you choose.

Classification modeling

3. Create a train/test split.
4. Build two classifiers that will produce predictions on the Current Condition of a load.
    a. Which of your classifiers is better and why? What does "better" mean in this context? (Select appropriate error metrics with justifications.)
    b. What features are most important in each classification model? Provide a table or plot and a brief description.
    c. Is there any reason we might be concerned with overfitting in these models? Do not re-fit your models, just provide a 1-3 sentence answer.

Context and critical reflection

5. What questions do you have about this data (list up to two)?

In 2-3 sentences, what steps would you take to address this gap in understanding if you were an employee?

Additional Guidance

Navigating Ambiguity

Even with the example above, we understand there is significant ambiguity in this question. The data scientist role requires a high degree of autonomy, so we would like to get an idea for how you approach a problem with relatively little guidance. The typical process involves back and forth with stakeholders. For this interview project, that is not something we would be able to provide fairly to all candidates. Being unable to ask those questions may require you to make assumptions. Use your best judgement, and please document those assumptions as you progress through the project.

Using your Resources

While we encourage collaboration at C.H. Robinson, this project is intended for us to get an idea of how you would solve this challenge, so do not ask others for help or search for solutions from other applicants. You will be asked to talk through your solution during the interview. Feel free to use other resources a professional data scientist might use.

Share the Project with Us

As a reminder, when you have completed this project, please submit the GitHub Repo and the link to your solution Submit your work through this form