Data Scientist Project

Description

Upon receiving the project file, you will have 48 hours to complete and send your submission. Please be sure to keep a copy of your submission and related material and be ready to speak to it in future interviews.

Data

Attached in the project file will be three CSVs containing records for Opportunities, Accounts, and Sales Representatives. Below is a description of the data and how the tables relate to one another.

Opportunity Table:

Each record in this table represents a closed sales cycle or 'Opportunity' in our Sales system. Our standard sales process proceeds through five stages (however they are not required to pass through each stage or in a sequential manner). Our sales stages are: Discovery (initial creation stage), Qualifying, Evaluation, Procurement/Negotiations, and Verbal Pending Close. Closed Opportunities are either 'Closed Won' or 'Lost'. We typically calculate the conversion rate (or probability an Opportunity will be 'Closed Won') as:

$$Conversion\ Rate\ = \frac{Total\ Records\ Closed\ Won}{(Total\ Records\ Closed\ Won + Total\ Records\ Lost)}$$

OPP_ID: Unique key for Opportunity records.
ACCOUNT_ID: ID representing the Account we are trying to sell to.
OWNER_REP_ID: ID representing the Sales Representative that owns that Opportunity.
OUTCOME: The outcome of the Opportunity – 'Closed Won' or 'Lost'.
CREATED_DATE: The date the Opportunity was created – Opportunities are in the first sales stage ('Discovery') when created.
CLOSE_DATE: The date the Opportunity was closed.
DATE_OF_QUALIFYING: A timestamp of the most recent time an Opportunity was moved into the second sales stage – 'Qualifying'.
DATE_OF_EVALUATION: A timestamp of the most recent time an Opportunity was moved into our third sales stage – 'Evaluation'.
DATE_OF_PROCUREMENT_NEGOTIATIONS: A timestamp of the most recent time an Opportunity was moved into our fourth sales stage – 'Procurement/Negotiations'.
DATE_OF_VERBAL_PENDING: A timestamp of the most recent time an Opportunity was moved into our fifth sales stage – 'Verbal Pending Close'.

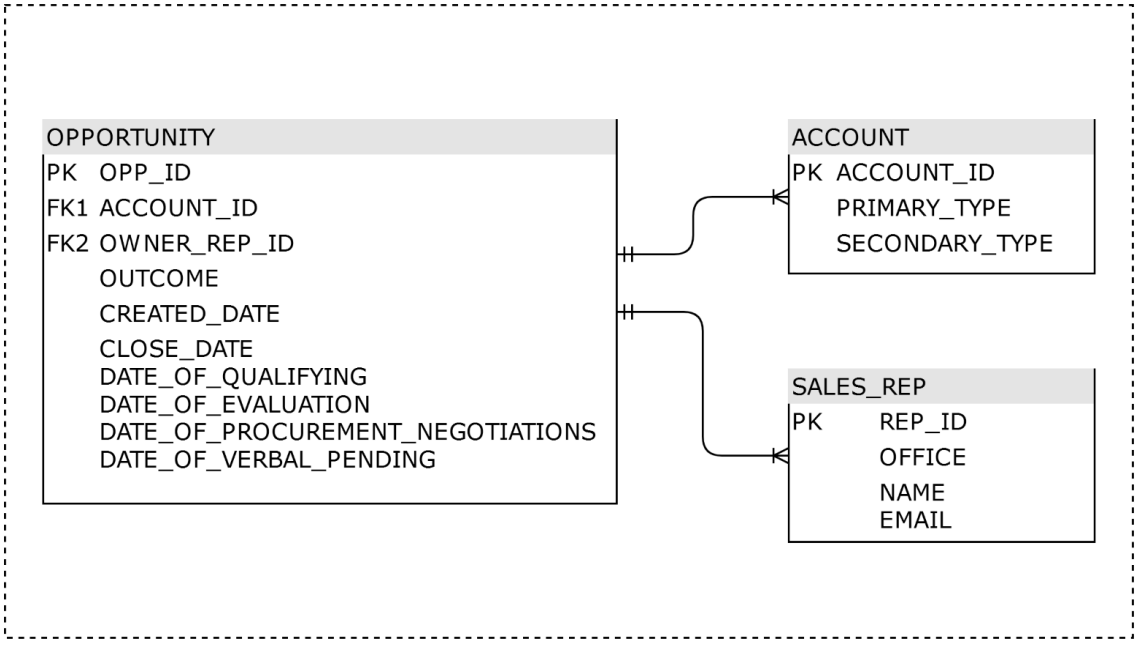
Sales Rep Table:
Each record in this table represents an individual responsible for selling PitchBook Subscriptions.

REP_ID: Unique key for Sales Representative records.
OFFICE: The regional PitchBook office the Sales Representative works in.
NAME: The name of the Sales Representative.
EMAIL: The email address of the Sales Representative.

Account Table:
Each record in this table represents an Account in our sales system.

ACCOUNT_ID: Unique key for Account Records.
PRIMARY_TYPE: The primary type of the Account.
SECONDARY_TYPE: The secondary type of the Account.

Entity Relationship Diagram – how the tables relate:



Questions for Analysis

Please use the data provided to effectively analyze and communicate your answers to the questions below. Feel free to use any data manipulation, analysis, and visualization tools that you are comfortable with. We encourage you to create any data visualizations that would support and communicate your responses – bear in mind we are also evaluating your analytical and data communication skills. Finally, clearly document your work and submit all associated files - SQL, Python, Excel Workbooks, Tableau Workbooks, Power BI Report etc.

1. Coding Skills:
    a. Write a Python/R script to clean and preprocess a dataset.
2. Statistics:
    a. Analyze given datasets to identify trends, outliers, and patterns.
3. Modeling Skills:
    a. Build a simple regression model to predict Sales won/lost based on account and sales rep profile. Please explain and justify your model choice, your train/test methodology, and explain its pros and cons with respect to the data set and problem.
4. Insights:
    a. From your answers to the questions above, what additional insights do you have about our business?