Data Science Challenge

1. **SQL (25 points)**

   Dataset: You are provided with a sample data table that has two columns: user_id (unique identifier of a customer) and transaction_date (the date which a customer made a transaction).

   Question: Please write SQL query (follow Presto SQL syntax) to produce a customer retention table (see example below).

   Raw data table

user_id	**transaction_date**
bdcff651-5a04-41e9-8c9a-83a5a192a420	2018-01-01
bdcff651-5a04-41e9-8c9a-83a5a192a420	2018-03-15
bdcff651-5a04-41e9-8c9a-83a5a192a420	2018-02-06
f763c01e-e46e-4be1-8dd6-a02da06269a8	2018-04-01
f763c01e-e46e-4be1-8dd6-a02da06269a8	2018-02-07
…...	…...

   Customer retention table

   The table should be interpreted in this way: 35% in the second row and second columns means there are 35% of customer who made their first transaction in the month of Jan (defined as “Jan Activation Cohort”) and subsequently made a transaction during the one month period followed their first transaction date.

	1st Month	2nd Month	3rd Month	…...
2018-01-01	35%	23%	15%	…...
2018-02-01	33%	26%	13%	…...
2018-03-01	36%	27%	12%	…...
…...	…...	…...	…...	…...

2. **Modeling (45 points)**

Dataset: You are provided with a sample dataset of a telecom company’s customers and a detailed explanation is as follows:

Column Name	**Column Type**	**Column Description**
State	String	The state where a customer comes from
Account length	Integer	Number of days a customer has been using services
Area code	Integer	The area where a customer comes from
Phone number	Alphanumeric	The phone number of a customer
International plan	String	The status of customer international plan
Voicemail plan	String	The status of customer voicemail plan
No. vmail msgs	Integer	Number of voicemail message sent by a customer
Total day minutes	Float	Total call minutes spent by a customer during day time
Total day calls	Integer	Total number of calls made by a customer during day time
Total day charge	Float	Total amount charged to a customer during day time
Total eve minutes	Float	Total call minutes spent by a customer during evening time
Total eve calls	Integer	Total number of calls made by a customer during evening time
Total eve charge	Float	Total amount charged to a customer during evening time
Total night minutes	Float	Total call minutes spent by a customer during night time
Total night calls	Integer	Total number of calls made by a customer during night time
Total night charge	Float	Total amount charged to a customer during night time
Total intl minutes	Float	Total international call minutes spent by a customer
Total intl calls	Integer	Total number of international calls made by a customer
Total int charge	Float	Total international call amount charged to a customer
Customer service calls	Integer	Total number of customer service calls made by a customer
Churn	Boolean	Whether a customer is churned or not

Question: (*Note:* You're free to use any programming language (Python, R, Julia) that you're familiar with and include the code together with your analysis.)

a) Perform exploratory analysis and extract insights from the dataset.
b) Split the dataset into train/test sets and explain your reasoning.
c) Build a predictive model to predict which customers are going to churn and discuss the reason why you choose a particular algorithm.
d) Establish metrics to evaluate model performance.
e) Discuss the potential issues with deploying the model into production.

3. Experiment Design (30 points)

Question: Following up on the telecom customer data, please write up a few points on how you plan to design an experiment to reduce customer churn with the outputs from the predictive model in the previous question.

You can include the following points but don't feel restricted in anyway.

a) Establish the primary objective of the experiment and create metrics for performance measurement.
b) Create null hypothesis and alternative hypothesis and discuss corresponding statistics.
c) Discuss how you setup the control and treatment group and overall experiment workflow.
d) Explain the risks of the experiment and how to mitigate the risks