Stori Data Science Challenge

Thank you for your interest in Stori. We estimate that this Data Challenge can take up to 2 business days to solve and submit results, if you have questions about the challenge don't hesitate to contact us and we will be happy to help you.

At Stori, you'll always be working with a diverse team, often including business analysts, engineers, product managers, and senior leaders. The ability to not just complete data science work, but also **to explain and present** your work is important. With that in mind, please deliver your answers as if you were handing off work to another data scientist joining your team. Tying together your intent, code, analysis and **results with visualizations** is the goal.

While we recognize sharing code is a common practice in the open source community, we ask that you refrain from sharing any part of this interview submission publicly to ensure a fair and equal interview process for all. We also ask that any code written by others is cited appropriately.

Challenge

For this challenge we will use a fake credit card dataset that is attached as *df.csv*, which includes information from a public Kaggle dataset with three added fields: *activated_date*, *last_payment_date* and *fraud*.

Question 1

1.1 Plot an histogram of the balance amount for all the customers.
1.2 Report any structure you find and any hypotheses you have about that structure.
1.3 Report mean and median balance, grouped by year and month of *activated_date*.

Question 2

2.1 Report in a table the following information for customers who activated their account and made their last payment during 2020: *cust_id* (excluding letters), *activated_date* (in format YYYY-MM), *last_payment_date* (in format YYYY-MM-DD), *cash_advance*, *credit_limit*, and a calculated field of *cash_advance* as a percentage of *credit_limit*.

Question 3

3.1 Build a predictive model for fraud.
3.2 What explanatory variable was the most powerful predictor for fraud?

We appreciate it if you include explanations of your intent, methods, conclusions and any assumptions. Please reply with the source code used to answer these questions in a clear and well-documented manner. If you attempted methods that didn't work or had ideas that could not be implemented, also include them in the final results.

Good luck!