Case Study Data Scientist - Support Data Products

Hello there, thanks again for your interest in Trade Republic. To kick off the interview process we have prepared a short technical challenge.

Business Case.

Trade Republic is the largest savings platform in Europe. We are an online broker with a banking licence and enable customers to save and build wealth through investing in capital markets.

We constantly evolve our product and have been expanding rapidly with new offerings such as interest, cards and bonds. As the product adapts, so do our support needs, and we are looking for great data scientists and engineers to join us in building data products that will help us provide intelligent, first class customer service 24/7.

One of the requirements of scaling support for millions of customers, means that we need to be able to help our agents triage incoming tickets through meaningful queues and prioritisation in real time. In order to achieve this, we use labels that we call “contact reasons” to classify incoming tickets and reflect why a customer is reaching out for help.

Your task

You’ve been tasked with helping your team develop a solution for classifying incoming support tickets. Our goal is to have a way to output the correct contact reason based off of the support ticket content.

Dataset

We’ve provided you with two mock datasets to work with.

1. The first dataset includes publicly available reviews of our app coming from the Apple Store, Google Store and TrustPilot reviews. **They are not labelled.**
2. The second is a list of customer service tickets with a label attached to it. **The label is the contact reason for the ticket.**

Tasks.

1. Using the *customer_reviews.csv* set, please provide and test data cleaning ideas. We are interested in your preprocessing skills and how you would do this at scale.

2. The *customer_tickets.csv* set has already been cleaned by your colleague. Use this dataset to classify each ticket with a contact reason label using an algorithm/-s of your choice. The list of labels can be found in the *contact_reason_labels.csv*.

Hint: You can use any available and ready-to-go solutions, i.e. packages/cleaning tools/embeddings/models etc.

Notes

- We want to understand how you approach the problem, feel free to add notes + comments.
- We recommend being time efficient here and want to respect your time. We expect you to spend no more than 4-8 hours of work on this task.

Deliverables.

- A reproducible notebook with your approach. It's important you show reasonable metrics to measure the performance of your approach, however it's not necessary to deploy the solution.
- A couple of slides or a one pager with the justification of your solution. If you make any hypotheses or want to challenge us, please develop.

Additional Questions

Here are some follow up questions which we would like to discuss with you during the case review interview.

- How would you deploy the solution?
- Would there be other strategies to improve prediction? What other data could be used to improve your solution?
- Which other data products would you build with these datasets?

How to submit your solution

Please zip your project and submit zip archive via the Greenhouse link attached to the email with the code challenge. Your dedicated recruiter will receive the notification about your submission and will send it to the team for review.