Machine Learning Engineer Take-home Project

BetterUp

Thank you again for your interest in the position! We are excited to dive into the ***fun*** part of the interview process where we focus on ***real world problem solving (and coding)*** instead of ***whiteboard logic puzzles.***

BetterUp

Your mission, if you choose to accept it...

Your marketing team wants to understand how Twitter users' behavior changes based on user age.

They've asked you to look at some tweet data and *tell some simple descriptive stories*, as well as see if you can come up with a way of *predicting user age* based on other characteristics.

BetterUp

Requirements: Part One

- **Build a statistical model to predict the age of users in *ages_test.csv***
  - You may use any of the information provided (profiles, friend networks, and mentions).
  - The training data is provided in ***`ages_train.csv`***
  - You may approach this task with age formulated either as a continuous or categorical/ordinal variable. Justify your choice in your report.
  - If you formulate age as a categorical/ordinal variable, create at least five category levels (e.g. age ranges) in your response variable.

- **Include some measure of user tweet sentiment as a predictor.**
  - It's up to you how to measure sentiment and what kind of feature to build from this measurement. Justify your choices and if it is useful as a predictor.

- **Include some measure of emoji use as a predictor.**
  - It's up to you what kind of feature to build from this measurement. Justify your choices and if it is useful as a predictor.

BetterUp

Requirements: Part Two

- **Devise a means of determining what “typical” tweet language looks like for a given group.**
  - Consider only the tweets in the training data written by users under the age of 30.
  - It’s up to you how to determine the group linguistic norm. As always, please justify your choices.

- **Identify the most unusual language, compared to the group norm.**
  - You may choose what you think of as a suitable metric to assess deviation from the group norm. Justify your choices.

* *If you run out of time to implement this solution in code but you have a strategy in mind, you may submit a written overview of your intended approach in lieu of a full solution. (Full implementations are preferred.)*

BetterUp

Outputs

Please submit a link to a GitHub repository or a single .zip file containing:

1. **Your code**
   - Submit your work in one or more Jupyter notebooks (.ipynb) and python source (.py)
   - Please use a recent version of Python 3 and your code should run.

2. Your predictions
   - Include a new CSV file named **`ages_pred.csv`** that includes two columns: ***`ID`*** and ***`Age`***.
   - The ***`Age`*** column will contain predicted ages for each Twitter user ID in the ID column.

3. Your findings
   - Your notebook(s) should tell a story. Please walk us through your methodology and findings. We value not just technical and intellectual acuity but the ability to communicate clearly and concisely to a non-technical audience.
   - Include answers to most or all of these questions:
     - Why did you choose the modeling strategy?
     - How well were you able to predict user age? What were the most important predictors?
     - What, if any, challenges did you encounter with data quality or data manipulation?
     - What, if any, are key limitations to your findings?
     - Would you feel confident in telling your company's marketing team that you can accurately target specific user ages with your model? Why or why not?
     - What's the key takeaway from this project that you'd give to a non-technical audience?
     - What would be the next steps to take the model into production?

BetterUp

Datasets

Filename	Description	Links
ages_train.csv	The training data, indicating the known age for each user ID in the set	download link
ages_test.csv	The test data (a set of Twitter user IDs for which the user's age is not known). The goal of the prediction task is to provide as accurate as possible a prediction of the ages of each user in this set.	download link
age_profiles.json	Twitter user profiles corresponding to the users in the training and test sets.	download link data format
age_tweets.json	Recent tweets from the users in the training and test sets.	download link data format
mentions.csv	A data set indicating users that have recently been mentioned by the users in the training/test set in tweets.	download link
mention_profiles.json	Twitter user profiles corresponding to the users mentioned in mentions.csv.	download link data format
friends.csv	A data set indicating users that the users in the training/test set are following	download link
friend_profiles.json	Twitter user profiles corresponding to the users in friends.csv	download link data format

BetterUp

Evaluation Criteria

Our values *guide* and *infuse* the work we do. They also provide a lens for evaluating this project. Some of the questions used to evaluate this project include:

- Is the project clear, concise, and understandable? #craftspersonship
- Does the project follow best practices? #craftspersonship
- Does the project feel alive and have a unique personality? #zest
- Did the candidate make reasonable trade-offs with respect to scope, quality, and time for the project? #craftspersonship
- Are the machine learning models validated appropriately? #craftspersonship
- Does the write-up communicate a clear narrative appropriate for expert and non-expert audiences? #empathy

** Learn more about our core company values at https://www.betterup.co/about*

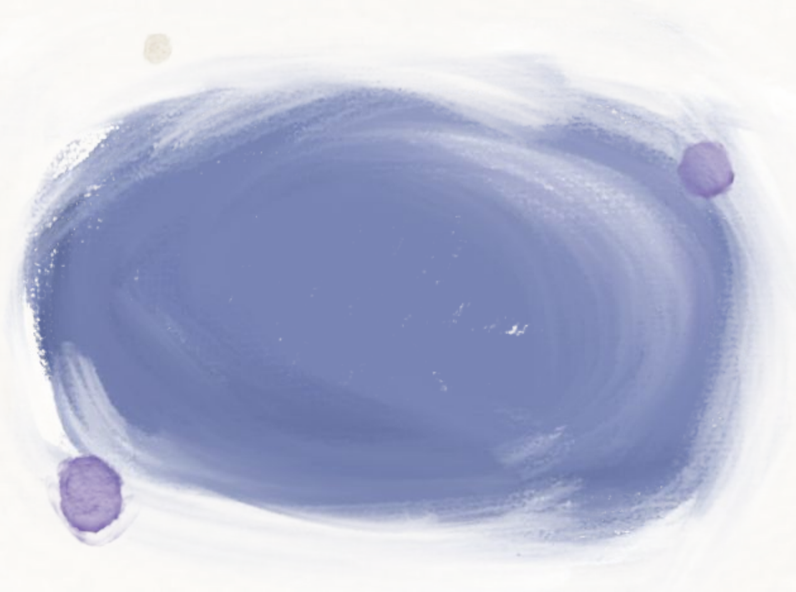
BetterUp

Project walkthrough and debrief

This project is intended to evaluate your approach to solving open-ended data science/ML problems **and** to provide a platform for further conversation with our team. After completion of your project, we will schedule a follow-up video call to walkthrough the code and hear from you how things went.

Some discussion topics include:

- How did you approach the project?
- What went well?
- What did you choose ***not*** to do?
- What would you change or improve?



Thank you, and reach out if you have any questions!

BetterUp