main | Go to file | Add file | Code

amandabgaeta Minor edits - re: intervals donut chart, grammar in presentation ... 20 seconds ago 19

data	Final fixes - Scatterplot titles and Getting rid of Gridlines	7 days ago
images	Corrected intervals instructors donut chart	7 minutes ago
notebooks	Minor edits - re: intervals donut chart, grammar in presentation	20 seconds ago
pdfs	Minor edits - re: intervals donut chart, grammar in presentation	20 seconds ago
.DS_Store	Py File, Viz files, and Updated Final Notebook	11 days ago
.gitattributes	Initial commit	12 days ago
.gitignore	Update .gitignore	12 days ago
LICENSE	Initial commit	12 days ago
README.md	Update README.md	9 days ago

README.md

Peloton-Class-Difficulty-Classification-With-Playlists

Introduction: Business Problem

About

No description, website, or topics provided.

Readme

GPL-3.0 License

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 99.9%
- **Python** 0.1%

With coronavirus and quarantine, companies in at-home fitness like Peloton experienced explosive growth. I was one of these people who invested in a Peloton bike and became a part of their online community. A common post is expression of excitement for their bike delivery coming soon and asking the group: “What class and/or instructor should I take first?”

A big part in selecting classes comes down to perceived difficulty based on the user. There isn’t a UI feature to filter by this, but at a class level you can see the difficulty level based on user ratings they are prompted for after completing a ride.

The final model has been constructed with these things in mind and classifies the perceived difficulty of the class based on Peloton class data including the Spotify features of the artists and tracks from the class playlists. As this was created with those receiving their Peloton bike in mind, the dataset and model only considers Cycling classes.

Access the final model code in /pdfs/Notebook.pdf OR the Jupyter Notebook version in /notebooks/3_models_with_writeups/1_Winning_Decision_Tree_manual.ipynb.

The Data

The dataset was custom made using a merging of:

- March 2021 Peloton class data extracted using the API and posted on Reddit by okaychristinakay (AGF_Peloton Classes March.xlsx)
- Kaggle datasets on Spotify artist and track data (kag_spot_artist_uris.zip, kag2_spot_artists.zip, tracks.zip) and
- Spotify artist and track data gathered via the Spotify search API using the Spotipy package.

Re-creation of the dataset can be accomplished by utilizing the files in the data/original_datasets folders (listed above) and running notebooks/0_create_dataset notebooks in order. This will get you to all dfs pickled and utilized in the final model.

Data Processing

Notebook.pdf includes cleansing, editing, functions, and feature engineering that was utilized in model iterations. Custom functions and pickle files needed to run the final notebook (1_Winning_Decision_Tree_manual.ipynb) are in /data/peloton_spotify_functions.py.

The /notebooks/1_modeling folder has variations of data analysis, model approaches, and notebook iterations that were used to get to the /notebooks/1_modeling and /notebooks/2_winning models and analysis were used to get to /notebooks/3_models_with_writeups.

Model Iteration

/notebooks/1_modeling/91_Modeling_start_manual is the notebook containing the iterative modeling process where nine different variations of data were tried across mostly untuned Decision Trees and Random Forest classifiers. Random Forest usage scored the best and is tuned in notebook 92_. After getting to a final Random Forest model, I recognized an interpretable model would be more advantageous for analysis; thus I went back and tuned the Decision Tree - making that the final product.

The model baseline accuracy from guessing the majority class was 50% (based on Intermediate classes) and was increased to 67-69% along with the f1-score, which was the goal metric.

Data Analysis

top features that drove difficulty classification in the final model included:

- whether a class is Low Impact
- how long the class is
- whether a class is an Intervals class were the

From the instructor perspective, Hannah Corbin was associated with Beginner rides and Robin Arzon and Olivia Amato were more likely to have Advanced rides.

Conclusion

Recommendations

In my initial search for class or instructor recommendations for first time riders, I found that there was a lack of information available. There were “top lists” of instructors with general recommendations, but all of them were different. This leads me to believe these are completely based on the publication and the writer's opinions. There was a lack of any data backed recommendations; thus I confirmed there is an opportunity in the market for a model like this to be implemented.

As this is an early iteration, this model could be made available using a simple service like Streamlit to test user engagement and satisfaction.

Future Work

Updated Peloton Class Data Ongoing via API

This model was created utilizing Peloton data from March 2021. As Peloton refreshes their classes a regular basis, it would be a priority to utilize the Peloton API in supporting this model ongoing. The evolution of the library may also bring revisitation to the model as the make up and audience of Peloton also evolves.

Peloton Playlist Data Formatting and Accuracy

In the Peloton class data, the song and artist list for a single class’ playlist is separate, which made it difficult to find the exact song. Zipping the two features together did not work as multiple artists can be listed for the same song (see song: “California Girls” artist: Katy Perry, Snoop Dogg). I made the assumption that popular songs users would recognize would be used on playlists; thus searching for the song title on Spotify will return the correct song data. This same assumption applies to searching and obtaining the artist level features. The next iteration should have songs paired with their artists for the most accuracy Spotify search results.

Potential Expanded Approach

Prior notebooks outside of the "final notebooks" folder document other paths I considered in creating a base dataset, especially in regards to the music related data. Editing in this approach could potentially impact the feature importance of music and playlist is for various levels of difficulty.

Model Improvement and Beta Testing

With each of the above considered, there is exponential opportunity to advance this work and model performance to the point where it can be implemented at Peloton and beta tested with users. As Peloton is very much carried by a positive customer experience overall, the first ride on the bike is a key moment in the users relationship with the brand and fitness overall.