Project 4

Spotify Skip Prediction

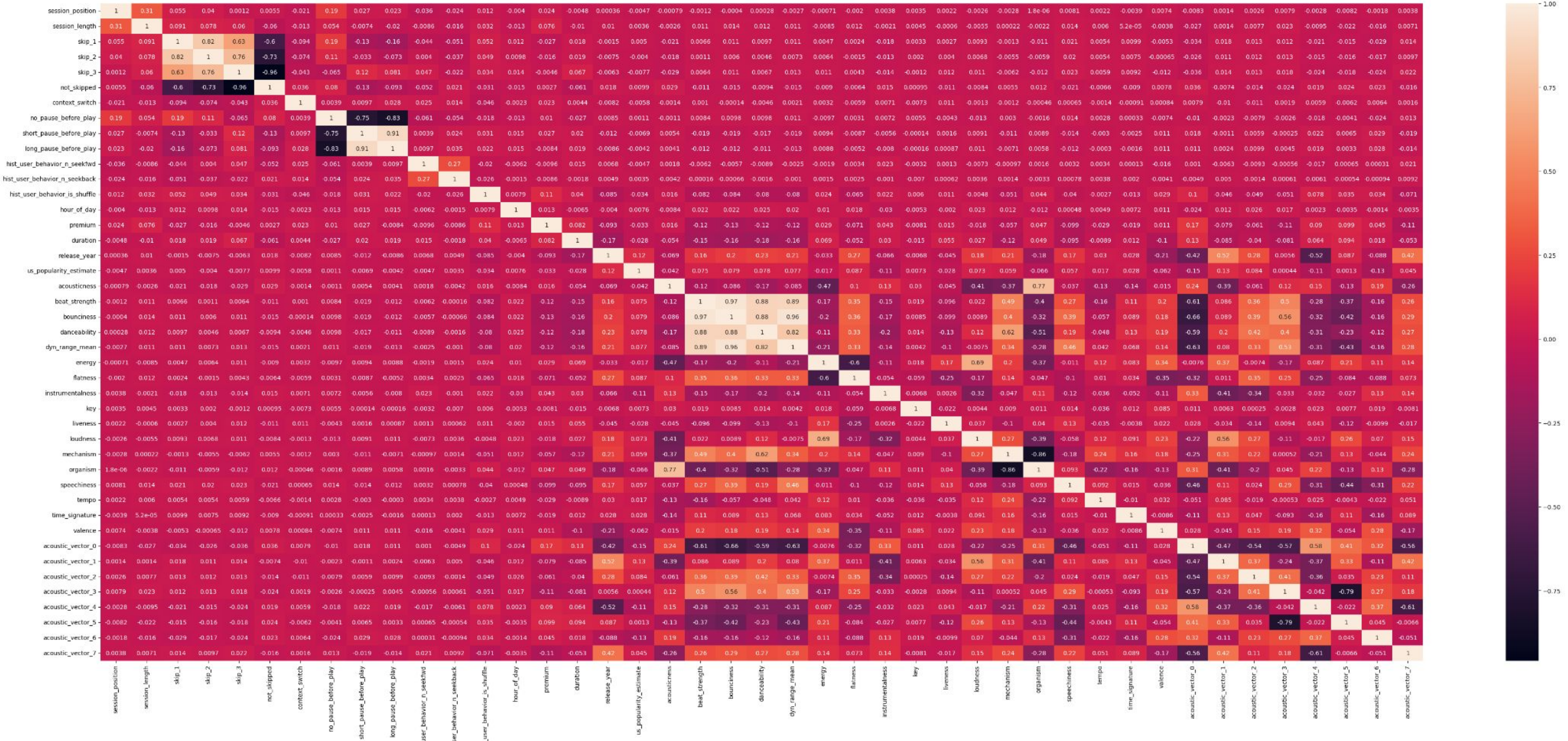
By Anika, Nanako, Peike

Dataset & Problem

- “Spotify Sequential Skip Prediction Challenge” by Spotify on AIcrowd
- 2 datasets: “track features” and “log”
- Each dataset contains:
  - track ID, session ID, and other identifiers
  - musical information
  - user behavior and history
  - Skipped or not?

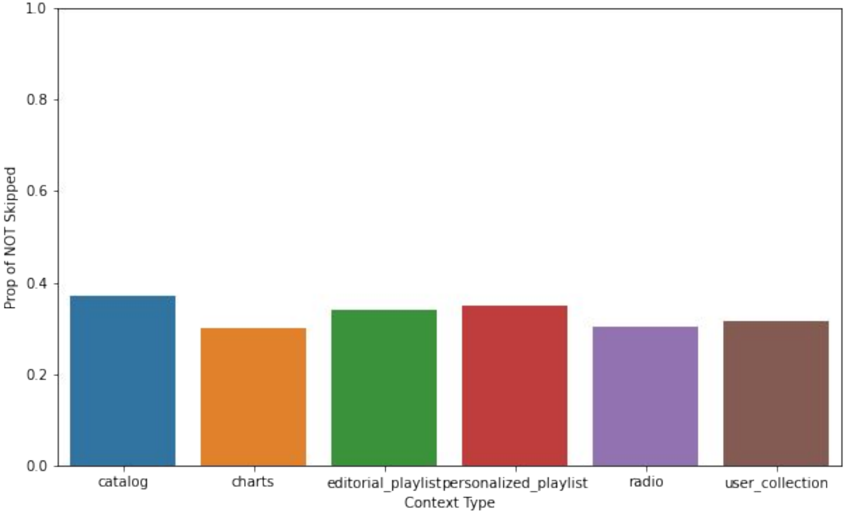
Challenge: Predict if a user will skip the current song

EDA - Heatmap of Correlation Coefficients



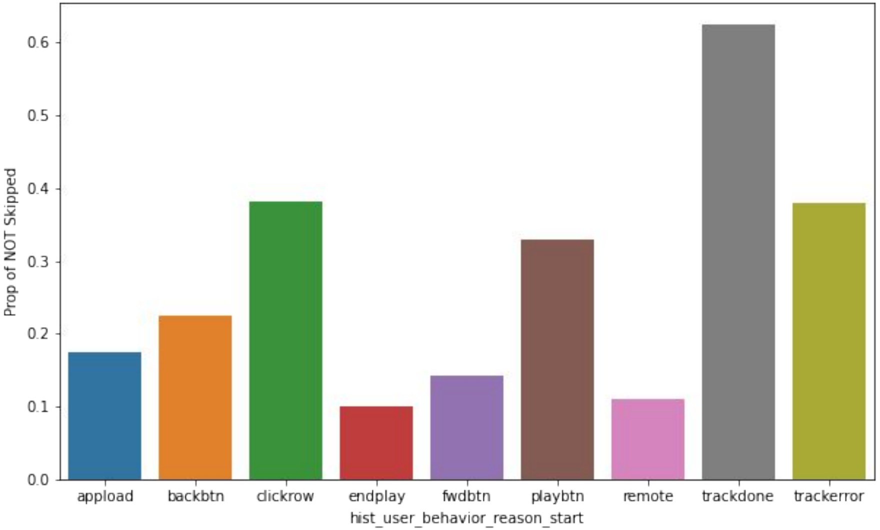
EDA - Examine Multi-class Categorical Variables

Proportion of Non-Skipped Songs by Context Types



Flat pattern indicates **weak predictor**

Proportion of Non-Skipped Songs by Play Reason



Sharp Contrast indicates **strong influencer**

Complex variables

- skip_1, skip_2, skip_3, not_skipped
  - Target variables that all depend on each other
- Categorical variables, e.g. “why the user started the song”, “why the user ended the song”
  - Possible values: “track done”, “track error”, “skip button”, etc.

Approaches for categorical variables:

a. Label encoding (hierarchical)

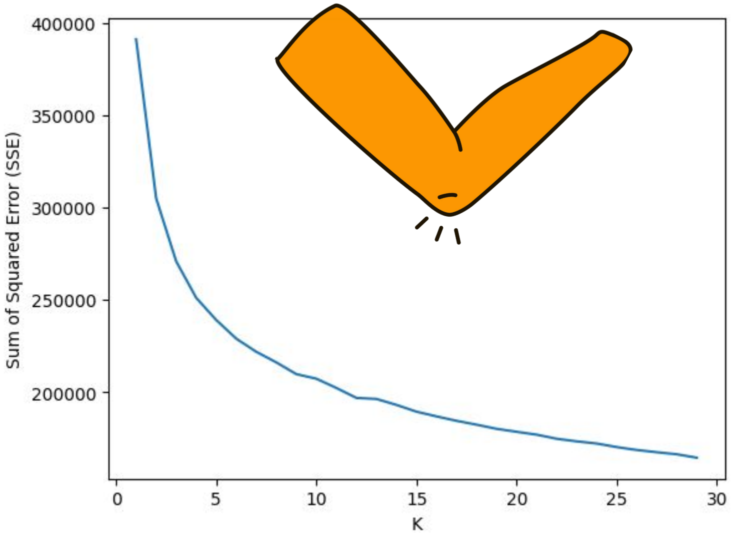
b. One-hot encoding (dummy encoding)

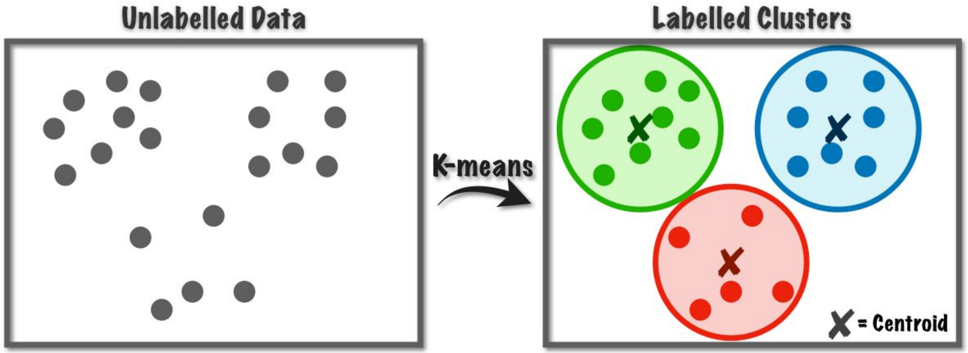
c. Exclude entirely

Why end song:		Why end song:
Track done	→	1
Track error	→	2
Skip button	→	3

Unsupervised learning: K-means Clustering

- 2 clusters: accuracy = 42%
- No sharp elbow
- 30 clusters: clustered based on instrumental features

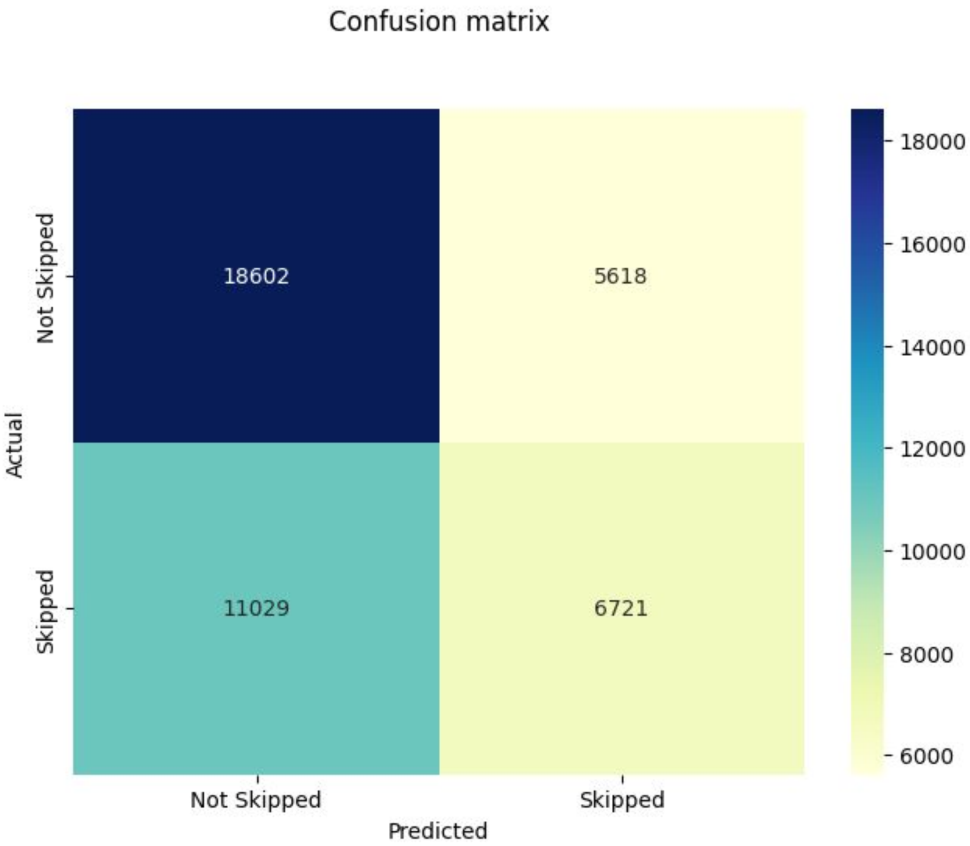




Logistic Regression: Two Approaches

- 1st model: Logistic regression using only numeric variables (no categorical variables)
- Confusion matrix displays results of all trials as a heatmap
    - ○ Ex) 18602 runs correctly predicted non skipped song as non skipped

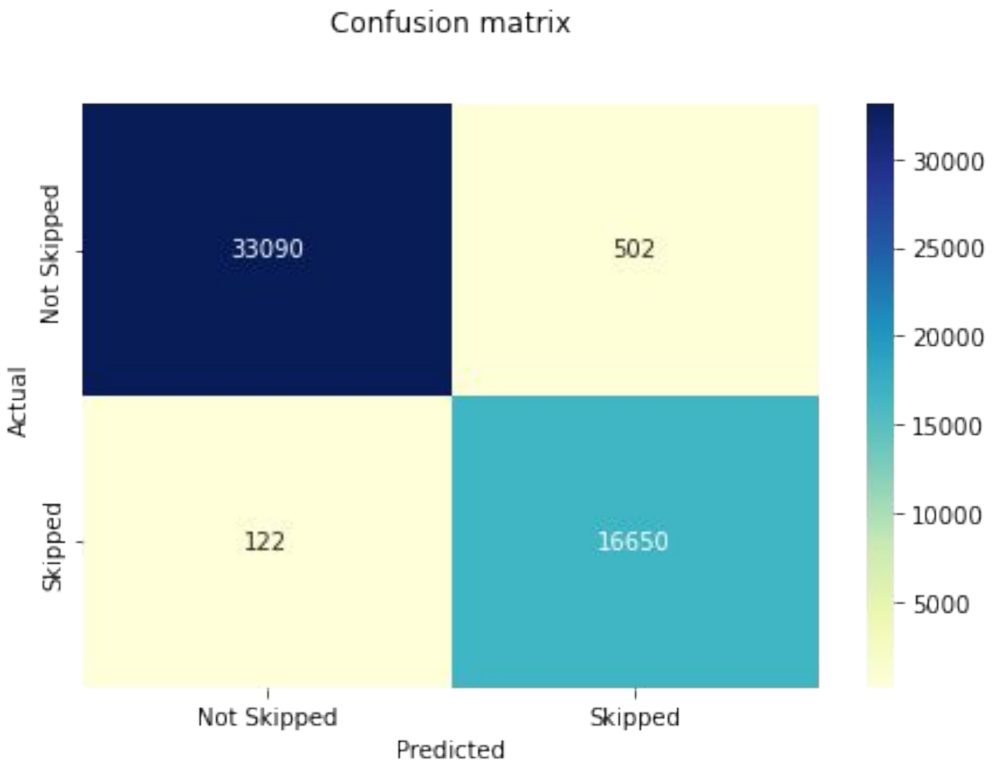
Accuracy score: 0.6034



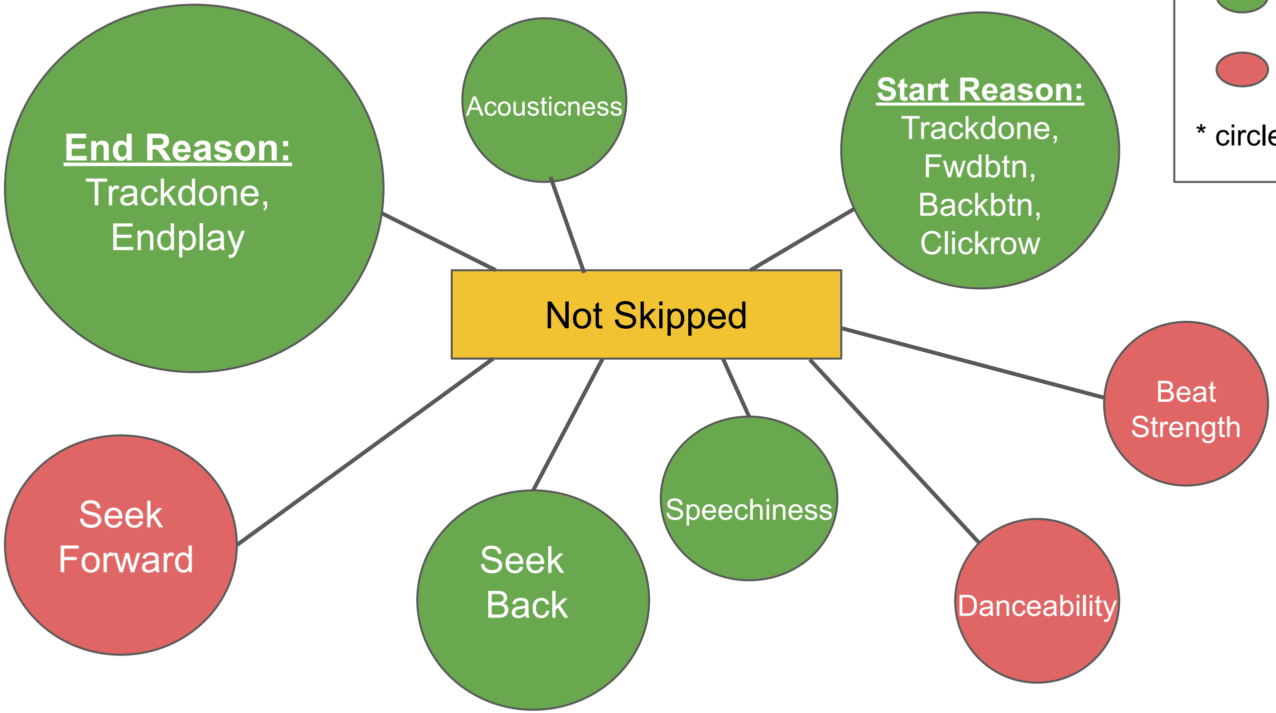
Logistic Regression: Two Approaches

- Change in feature selection: dummy encoding the “strong influencers” and remove the “weak predictors”
- **Prediction Accuracy** on test set:

98.8%



Top Important Features based on Coefficients



End Reason:
Trackdone,
Endplay

Acousticness

Start Reason:
Trackdone,
Fwdbtn,
Backbtn,
Clickrow

Not Skipped

Beat Strength

Seek Forward

Seek Back

Speechiness

Danceability

🟢 positively associated

🔴 negatively associated

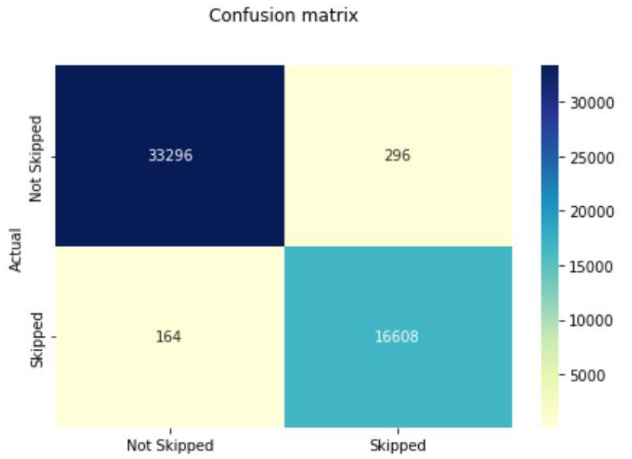
* circle size represents strength

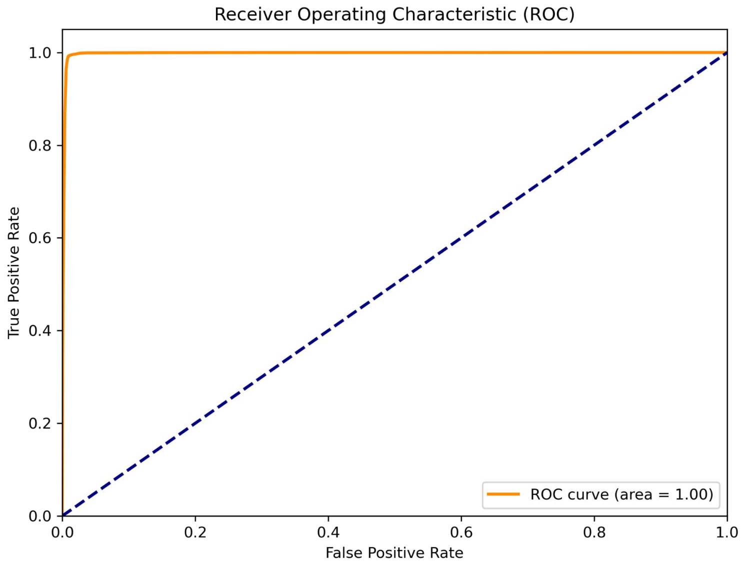
Random forest - Better Predictions

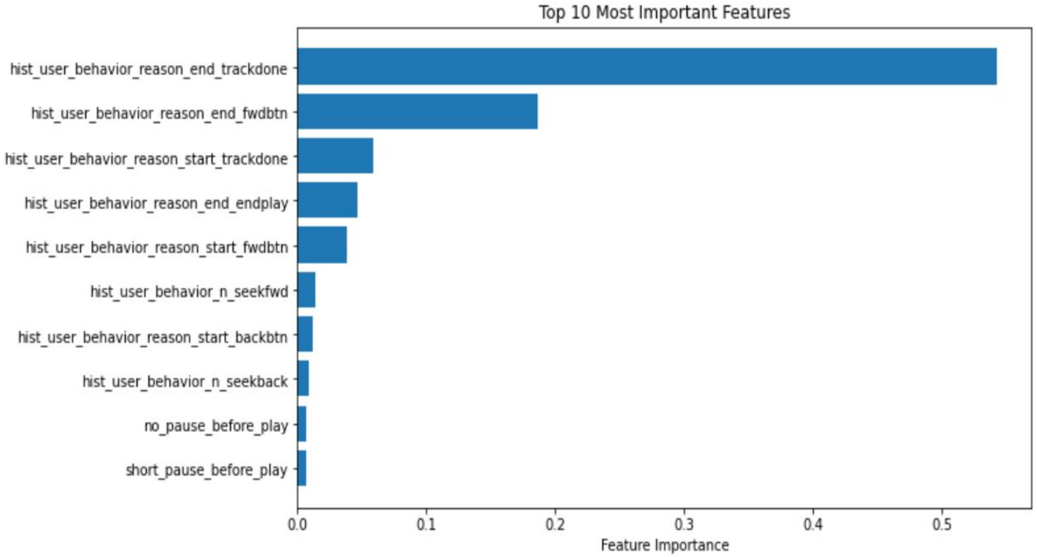
- Applied random forest classification with 100 estimators
- Model has strong ability to classify not-skipped vs. skipped — ROC curve

99.1%

prediction Accuracy on test set







Conclusion

- 3 Machine Learning Techniques:
  - Logistic Regression w/ & w/o categorical variables
  - Random Forest
  - Unsupervised Learning
- Logistic Regression Model with selected categorical variables had the highest accuracy: 0.9908
  - Including lowly correlated variables created a stronger model than only including strongly correlated variables
- Better understanding of the data collected by Spotify can help build stronger models
  - Ex) Applying non-linear relationships between variables

Thank you!

Links and sources

Headphones image: https://www.publicdomainpictures.net/en/view-image.php?image=212667&picture=black-headphones

K-means diagram: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c