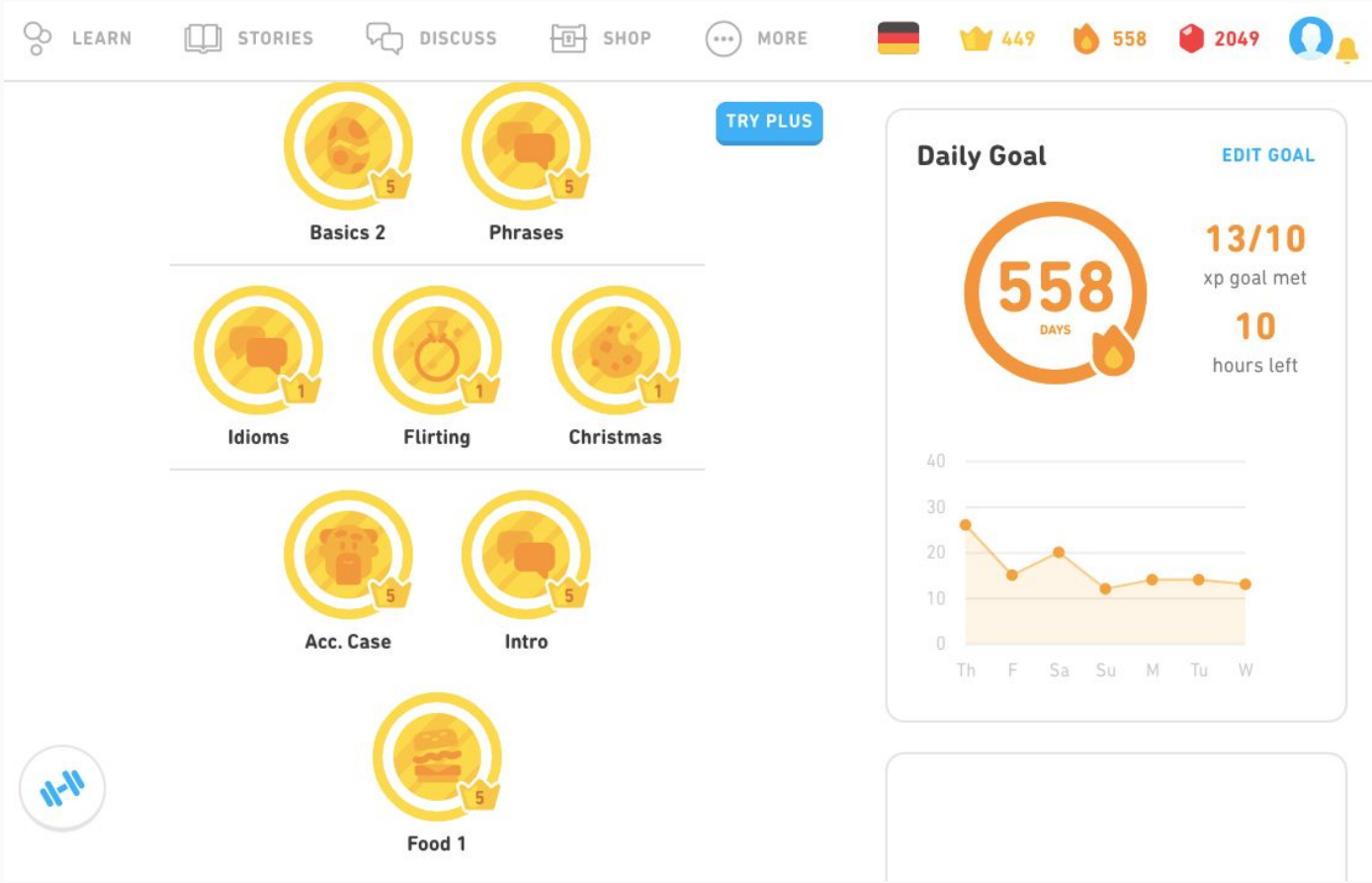
Second Language Acquisition Modeling

An analysis of data from Duolingo

By Kate Christensen

What is Duolingo?

Home page for a course a user is enrolled in:



About the Data

- Each observation is an individual word a user encounters within an exercise
- The particular dataset analyzed was Spanish speakers learning English
- This dataset had over 2.6 million observations

Data Features

- **User data:**
  - Unique id for each user
  - Countries user has used the app in
- **Exercise data:**
  - Exercise format
  - Exercise index
  - Response time
    - Response times longer than 60 seconds were cut
  - Prompt used
  - Session type
  - Client type
- **Token data:**
  - What the word was
  - Part of speech
  - Morphological info
    - Plural, definite, possessive, etc.
  - Syntactic info
    - Not included because previous work on this data indicated a lack of confidence in its accuracy

10 IN A ROW

Translate this sentence

I do not understand the Japanese class at all. I am in trouble.

日本語 の 授業 が 全然

分かり ませ ん こまり ます

つまらなかった

つまらない 困り

こまって

You are correct

CONTINUE

Type what you hear

Je peux être un mauvais garçon

Translation:

I can be a bad boy.

CONTINUE

Translate this sentence

Cat.

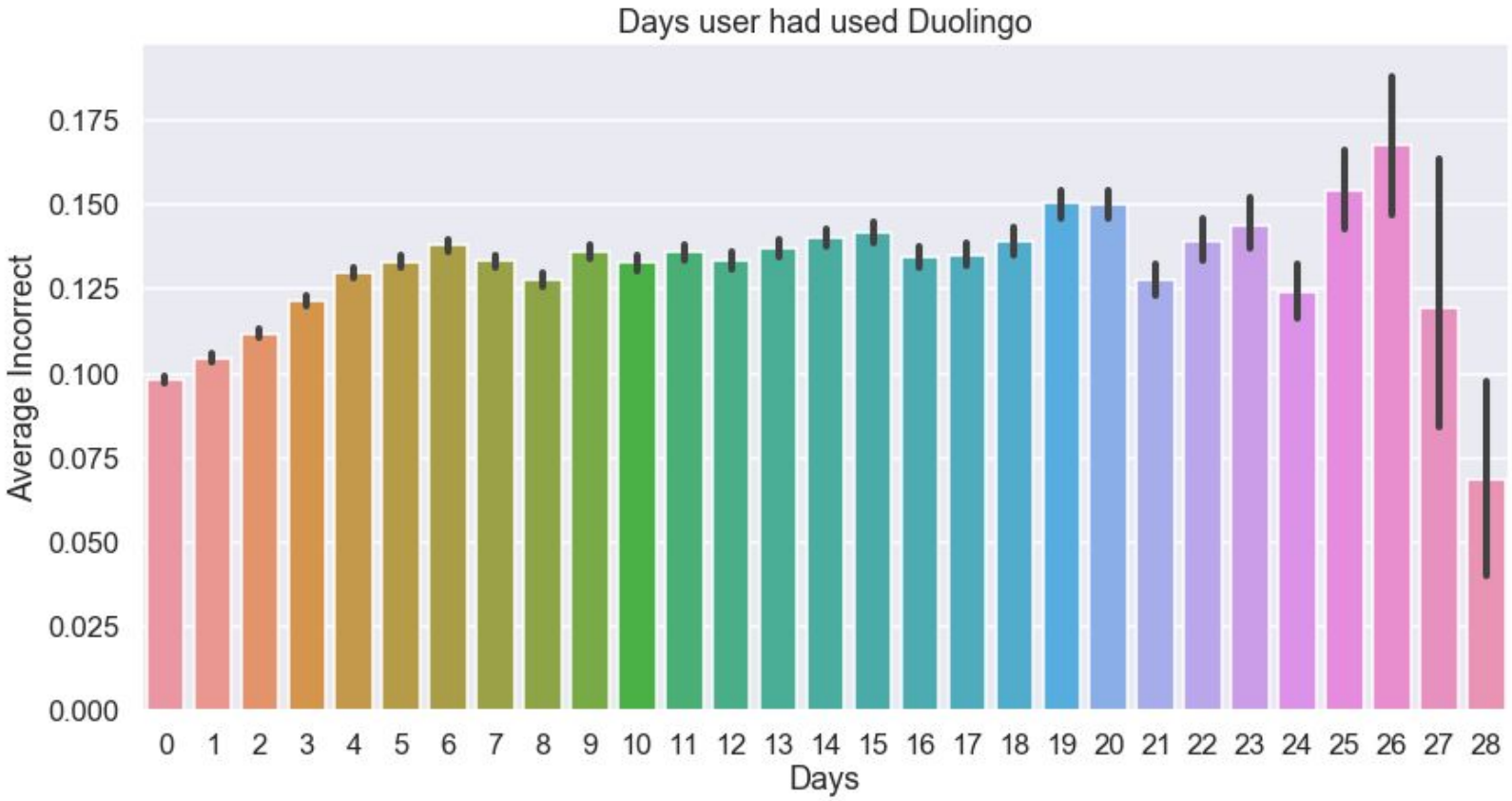
Cat

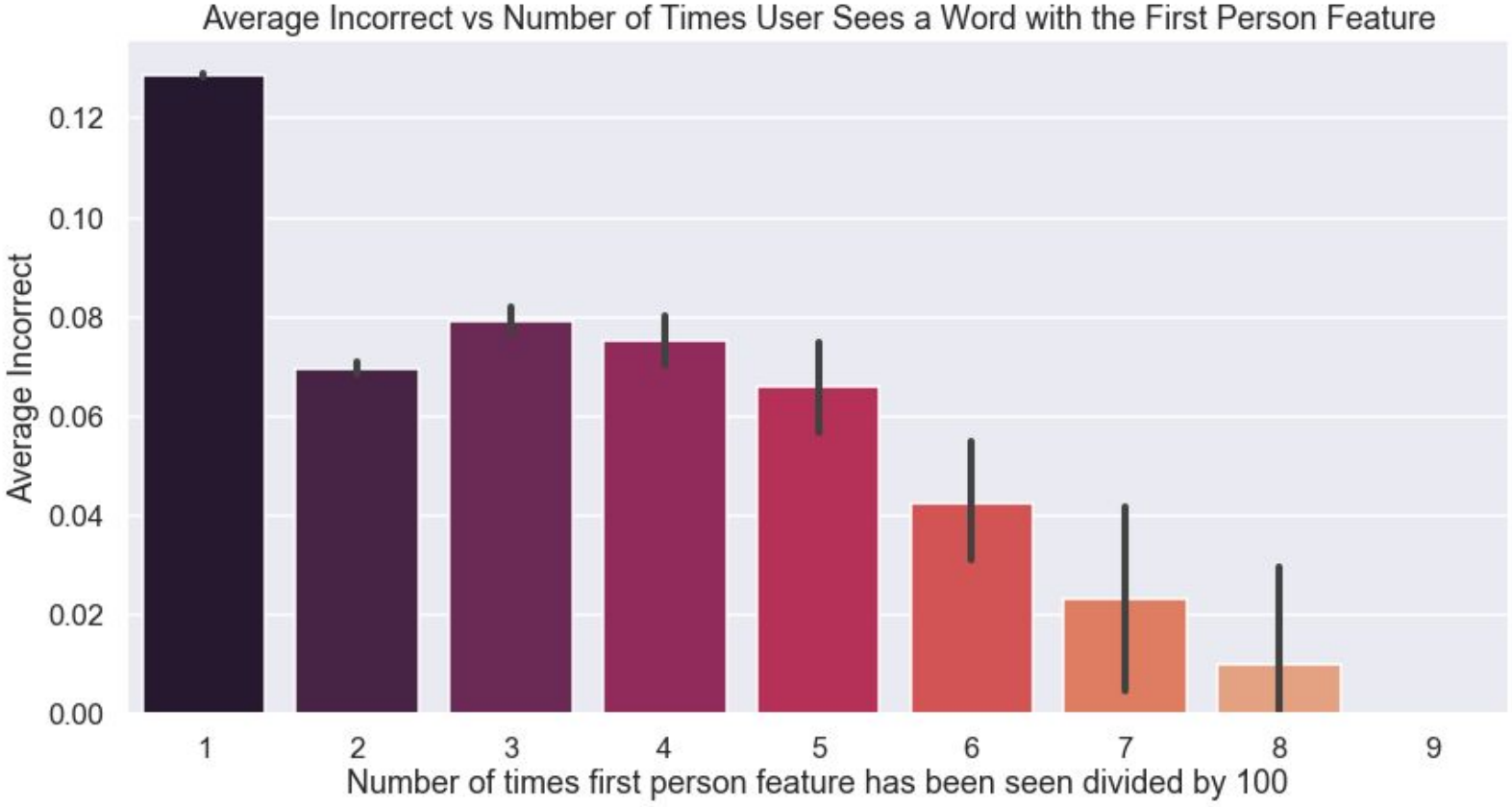
You are correct

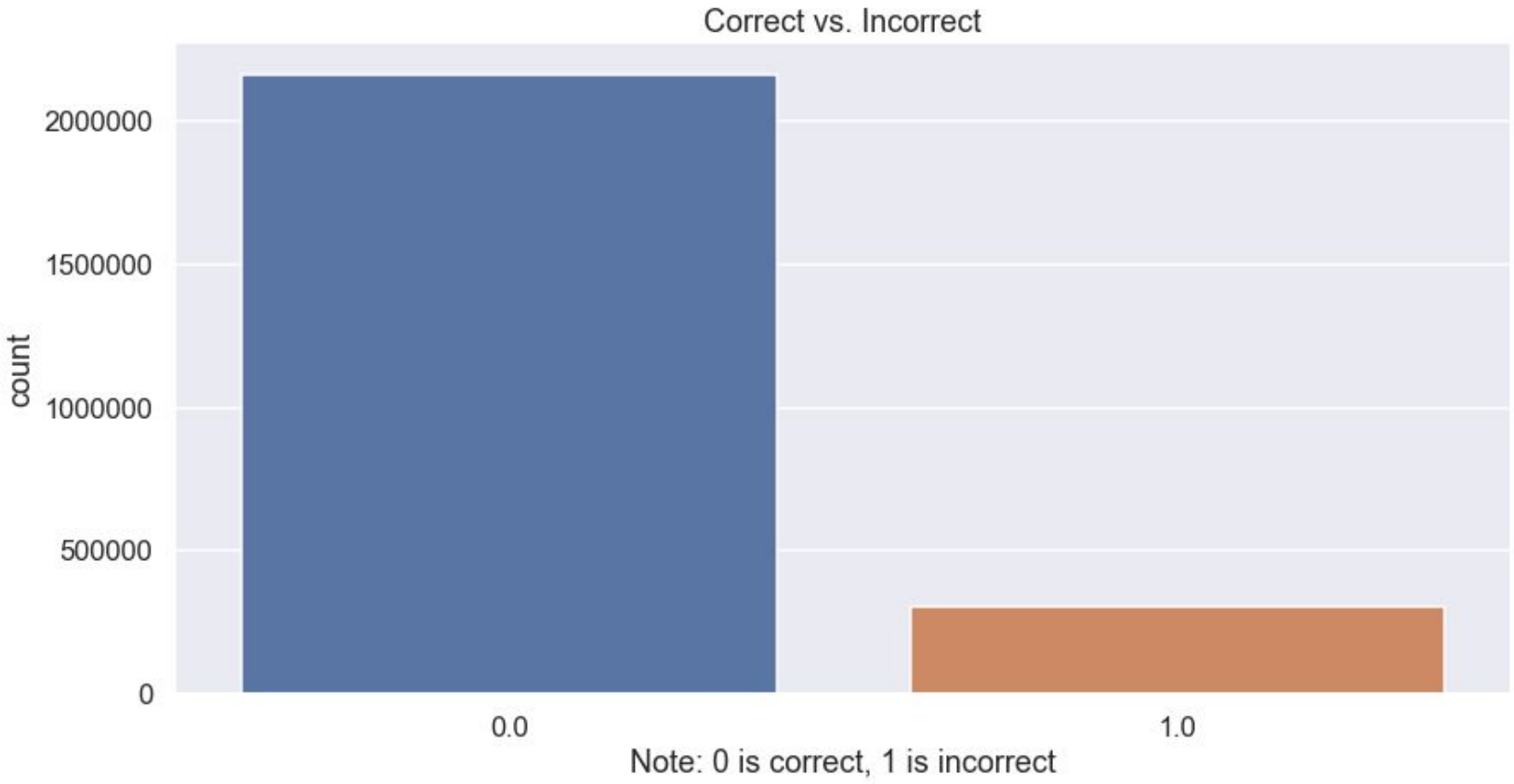
CONTINUE

Feature Engineering

- English:
  - Is English (the target language) a *de facto* language in at least one of the countries the user has used the app in?
- Morphological features and part of speech:
  - Is this the user's first, second, third, etc. time seeing this particular feature or part of speech


Days user had used Duolingo
Average Incorrect
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Days


Average Incorrect vs Number of Times User Sees a Word with the First Person Feature
Average Incorrect
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
6
7
8
9
Number of times first person feature has been seen divided by 100

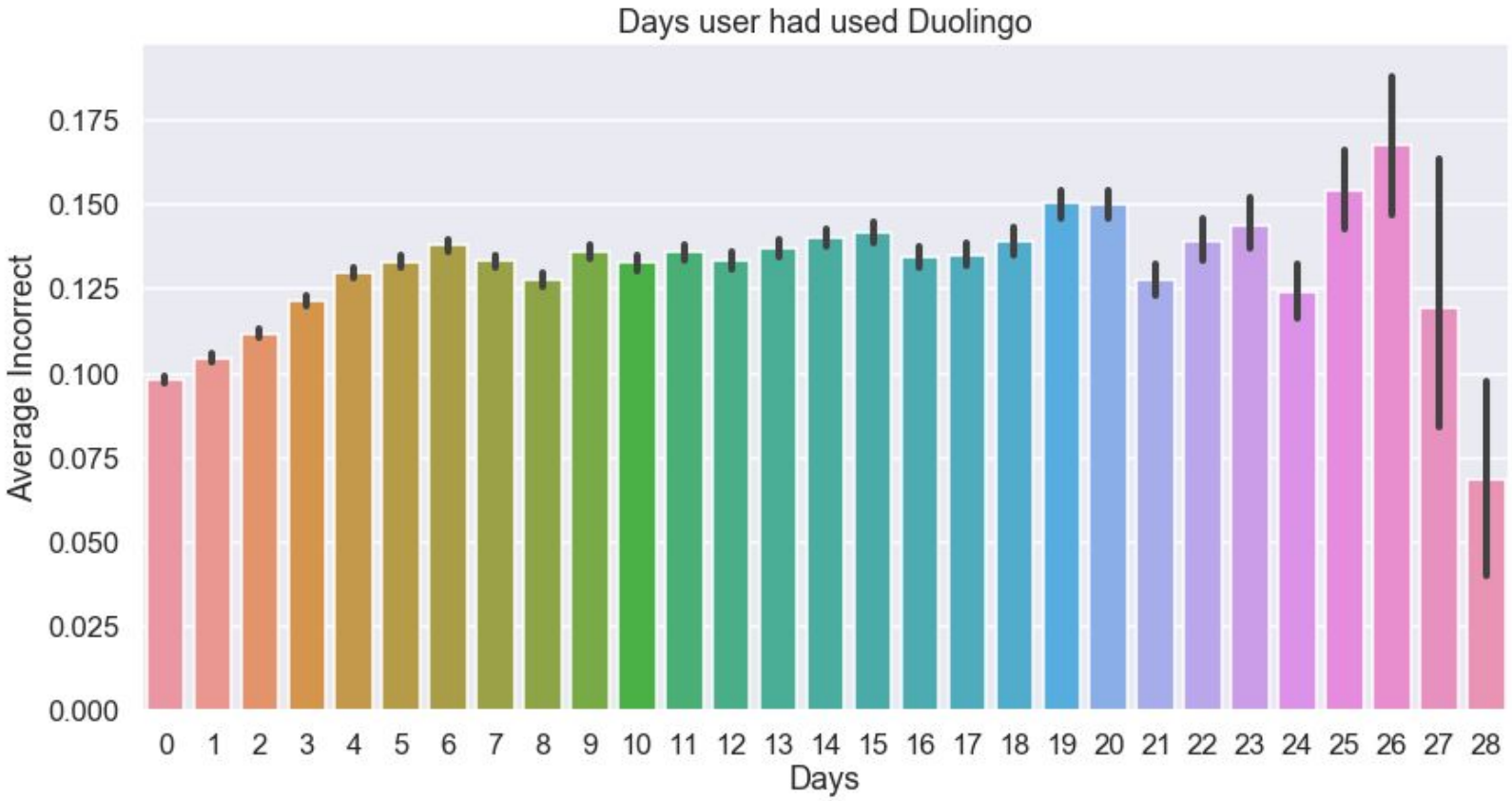

Correct vs. Incorrect
2000000
1500000
1000000
500000
0
count
0.0
1.0
Note: 0 is correct, 1 is incorrect

Addressing potential issues:

- Dimensionality
  - Final data frame has 58 features
  - Principal Component Analysis
  - Maintain 85% of variance: 36 features
- Imbalanced Data
  - Oversampling: SMOTE (Synthetic minority oversampling technology)
  - This greatly increased the F1 value (0.002 -> 0.29)

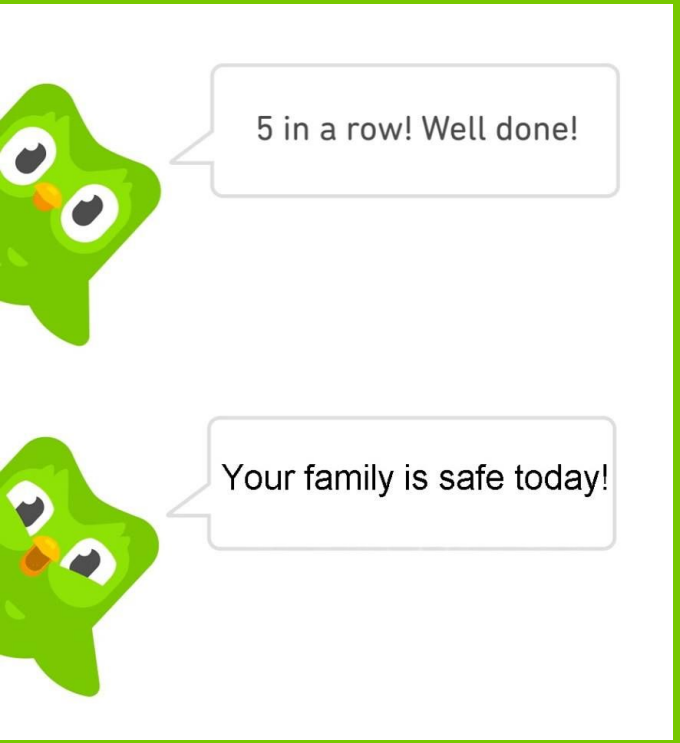
Results:

- Utilized Random Forest Model
- Accuracy: correctly predicted observations/total observations (0.53)
- F1: keeps a balance between Precision and Recall. (0.29)
  - Useful for data with uneven class distribution
  - Model has a good recall (0.776), but is imprecise (0.18)
  - Selects a lot of relevant items, but not all relevant items are selected
- AUC ROC: False Positives vs True Positives (0.64)


Days user had used Duolingo
Average Incorrect
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Days

Conclusion

- Modeling can help developers maintain a consistent difficulty level
  - Maintaining this difficulty level can help with user retention
- Users, primarily at the beginning of using the app, can be less predictable
  - Collect data for users who have been using the app for longer
- Whether or not the user was in an English speaking country seemed to be important
  - Tailor user content by location



Thank you!

```
[123]: forest = RandomForestClassifier(n_estimators=100, criterion='gini', max_depth= 5, class_weight="balanced")

[124]: fit_model_and_print_metrics(stk_indices, X_res, y_res, forest, print_metrics=True)

       accuracy_score: 0.7623190628776095
       precision_score: 0.7406164393044747
       recall_score: 0.8048450742400465
       f1_score: 0.7685893624664137
       roc_auc_score: 0.7623190628776095

[124]: {'accuracy': 0.7623190628776095,
        'precision': 0.7406164393044747,
        'recall': 0.8048450742400465,
        'f1': 0.7685893624664137,
        'roc_auc': 0.7623190628776095}

[125]: fit_model_and_print_metrics(stk_indices_test, X_te, y_te, forest, print_metrics=True)

       accuracy_score: 0.5349370472558669
       precision_score: 0.18063739000198628
       recall_score: 0.7761605271382962
       f1_score: 0.29306054132113857
       roc_auc_score: 0.638447716909507

[125]: {'accuracy': 0.5349370472558669,
        'precision': 0.18063739000198628,
        'recall': 0.7761605271382962,
        'f1': 0.29306054132113857,
        'roc_auc': 0.638447716909507}
```