

Personality Identification

...

Nicholas Morris

Machine Learning

Wrangling

Data Handling

Removing unnecessary columns. Joining the data sources together. Filling in missing values with None.

Feature Engineering

Transformations

Converting categorical columns to binary data points.

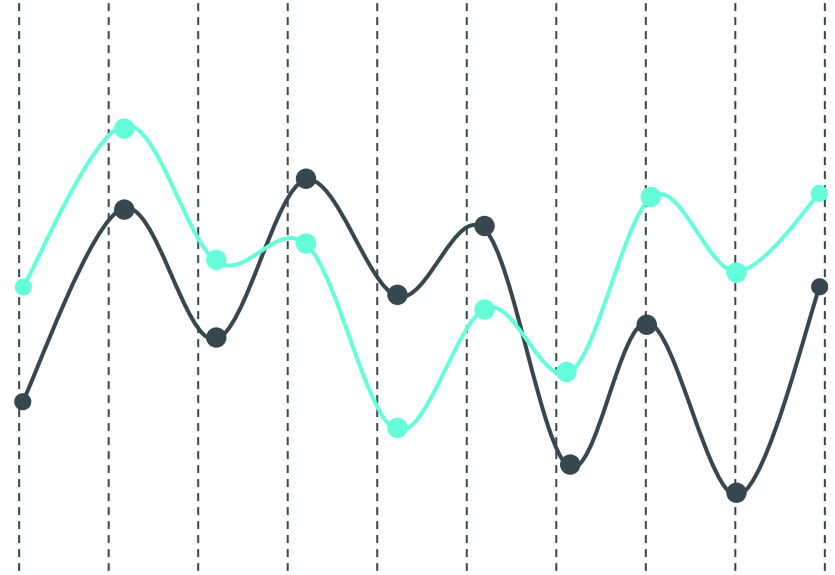
Attempting feature transformations such as Atwood Numbers, Binning, Reciprocals, and Interactions.

Modeling

Predictions

Training logistic regression, XGBoost, and deep learning neural network models.
Evaluating performance.
Computing feature drift to signal retraining.


Wrangling



—

Dataset

Below is the first user in the data. There are 8,328 total users and 26 columns. The target we are predicting is mbti_personality. [\[Link to the dataset and code\]](#)

id	mbti_personality	description	verified	followers_count	friends_count	listed_count	favorites_count	statuses_count	number_of_quoted_statuses	...	total_media_count	number_of_tweets_scraped	average_tweet_length	average_retweet_count	average_favorite_count	average_hashtag_count	average_url_count	average_mentions_count	average_media_count	tweet
16081623	infp	 {INFP}...	0	1904	782	67	133836	410600	14	...	114	200.0	11.785	3003.580	0.980	0.25	0.185	0.695	0.57	@andresit...

Removing Unnecessary Columns

`id_str`, `name`, `screen_name`, and `location` were removed from the dataset.

Joining Data

The data came in three separate parts:

1. MBTI Personalities Of The Users
2. User Information & Statistics
3. 200 Tweets Per User

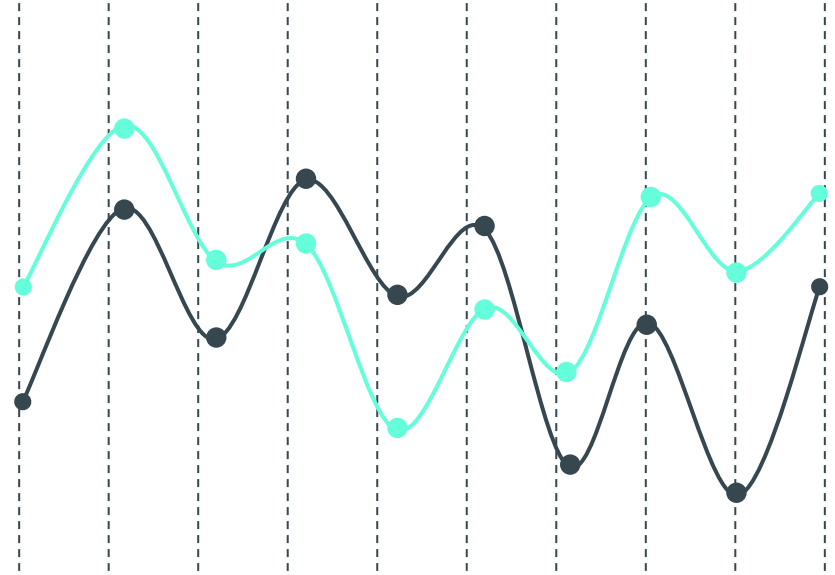
The data was joined together on the id column, and all of the tweets were merged together into one column.

The kernel kept crashing when processing the tweets, so, they were left out of the final model.

Missing Values

Missing values in the tweets and description were replaced with None.

Feature Engineering



—

Binary Data

verified was converted to binary data points.

Text Data

description was converted to binary data points for every instance of a word. The positivity of description was also computed.

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:

$$(x - y) / (x + y)$$

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as $1 / x$.

Reciprocals did not improve model performance, so, it was left out of the final model.

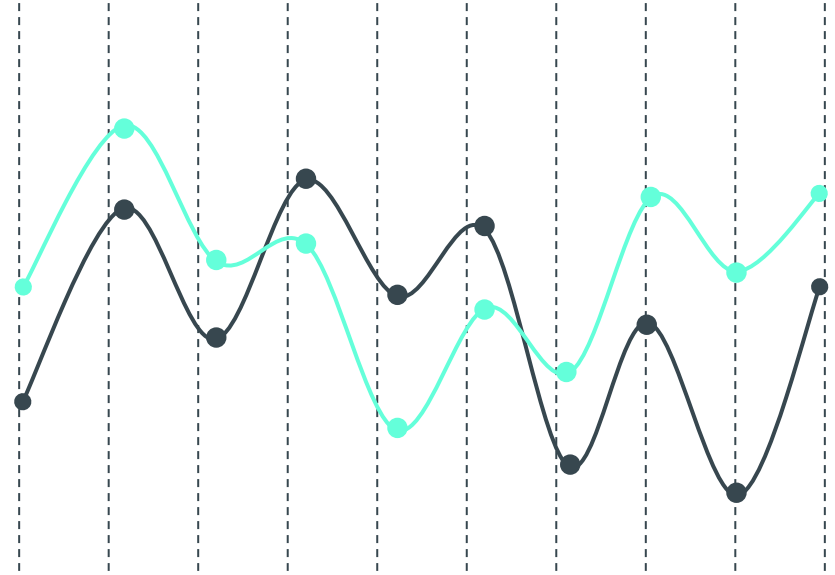
Interactions

An interaction is when two variables x and y are calculated as $x * y$.

Reciprocals were fed into this calculation to generate x / y as well.

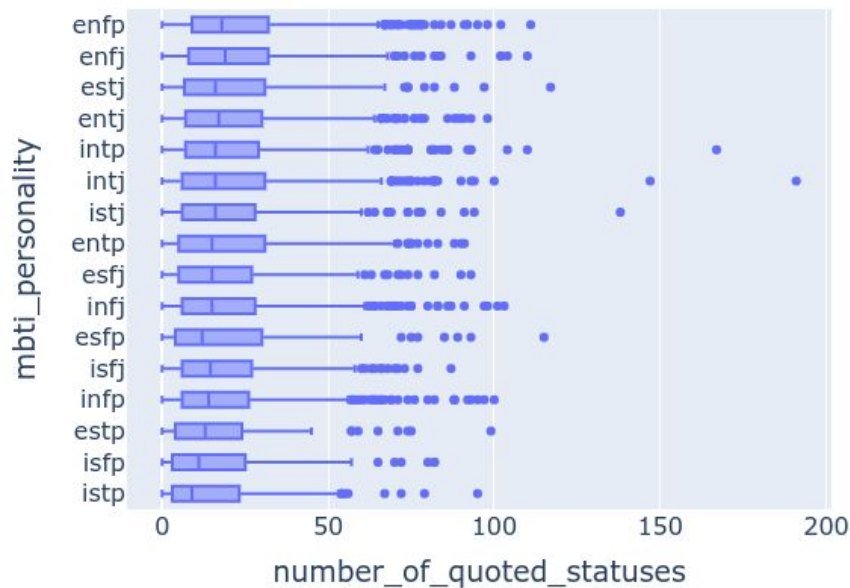
Interactions did not improve model performance, so, it was left out of the final model.

Data Exploration



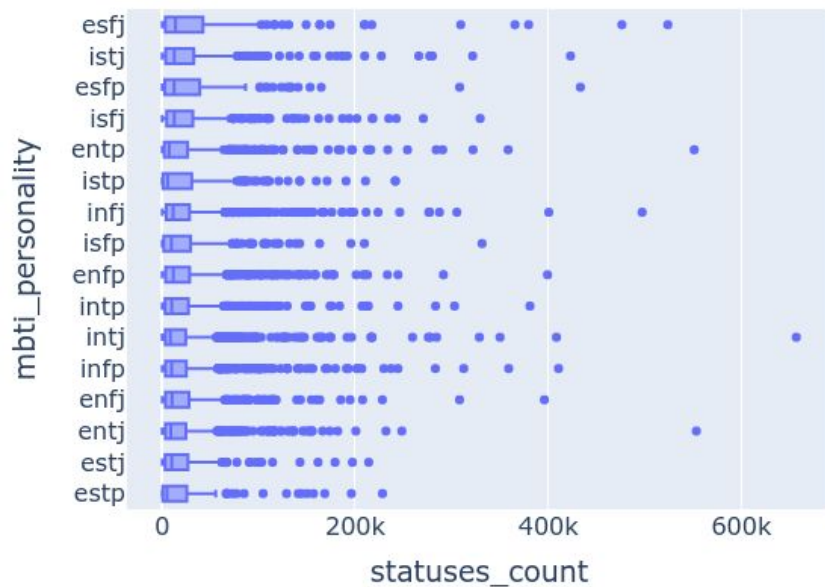
—

number_of_quoted_statuses vs. mbti_personality



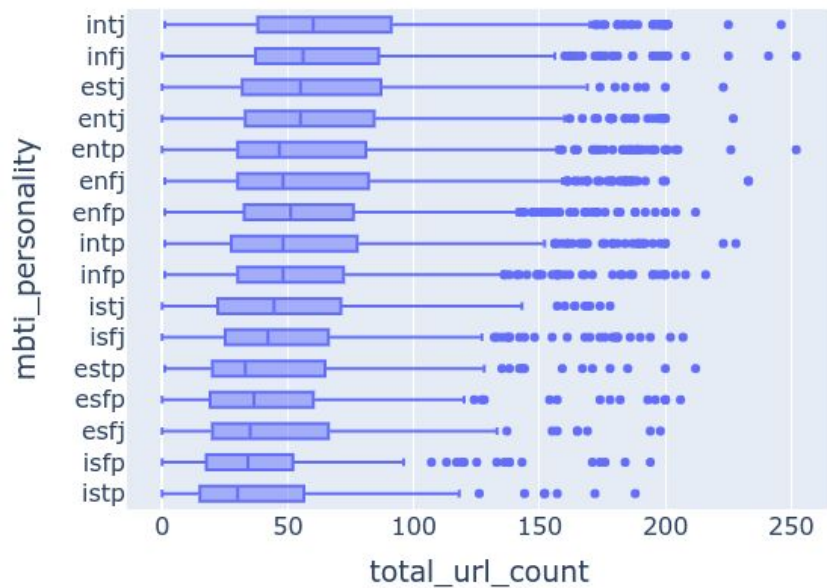
We can see that the number of quoted statuses varies slightly by personality. ENFP has a median of 18 quoted statuses, and ISTP has a median of 9 quoted statuses.

statuses_count vs. mbti_personality



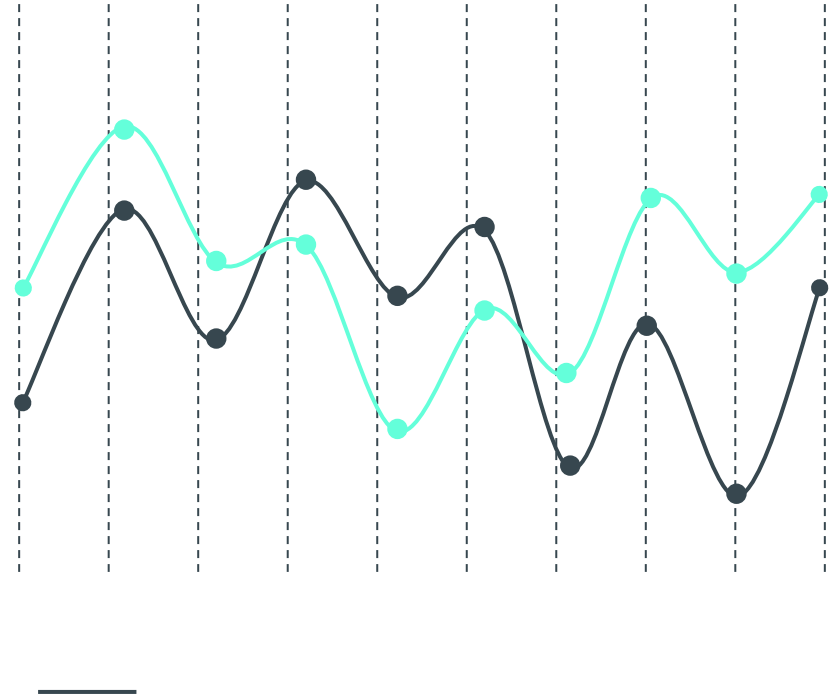
We can see that the statuses count varies slightly by personality. ESFJ has a median of 13.8 thousand statuses, and ESTP has a median of 4916 statuses.

total_url_count vs. mbti_personality



We can see that the url count varies slightly by personality. INTJ has a median of 60 urls, and ISTP has a median of 30 urls.

Modeling



Model Parameters

Logistic Regression

Library: scikit-learn
Penalty: L1
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: $1e-4$
Max Iterations: 100

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
 0.001, 0.01, 0.1
Max Depth:
 5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
 0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
 32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Logistic Regression

Accuracy: 0.83

F1: 0.79

In Control: 98.2%

Model Indicators:

1. number_of_quoted_statuses
2. favourites_count
3. average_tweet_length
4. description_esfp
5. description_intj

XGBoost

Accuracy: 0.85

F1: 0.83

In Control: 98.8%

Model Indicators:

1. description_intj
2. description_in fj
3. description_enfp
4. description_infp
5. description_enfj

Neural Network

Accuracy: 0.17

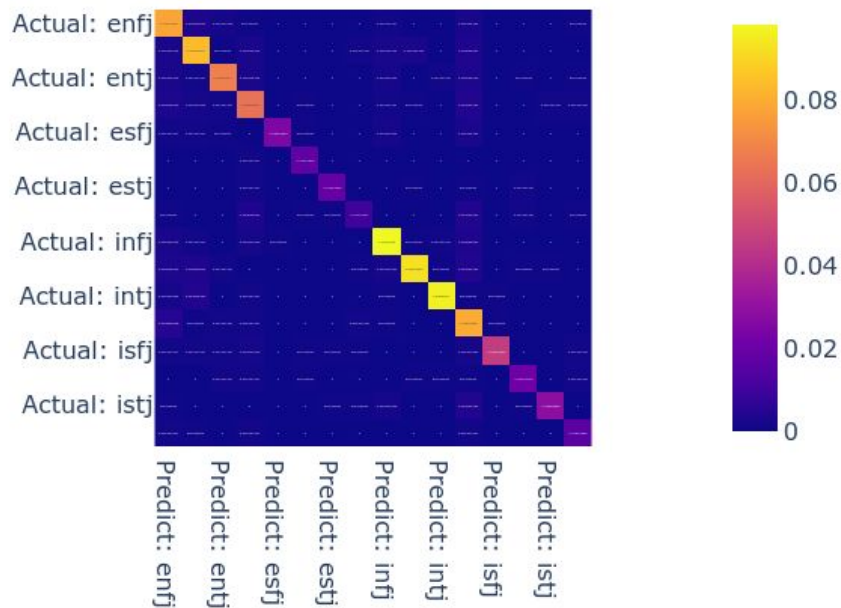
F1: 0.11

In Control: 100%

Model Indicators:

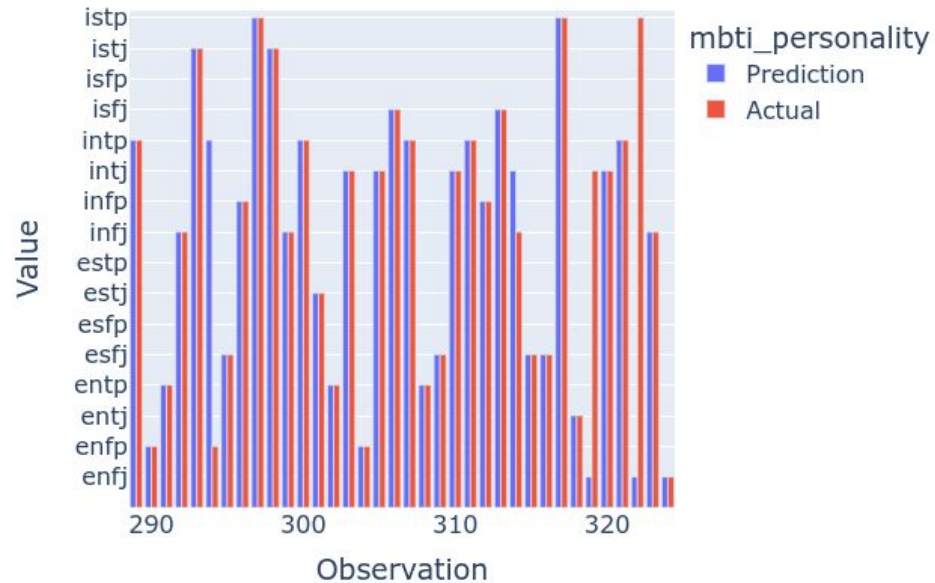
1. total_retweet_count
2. total_media_count
3. average_media_count
4. total_mentions_count
5. description_positivity1

Confusion Matrix

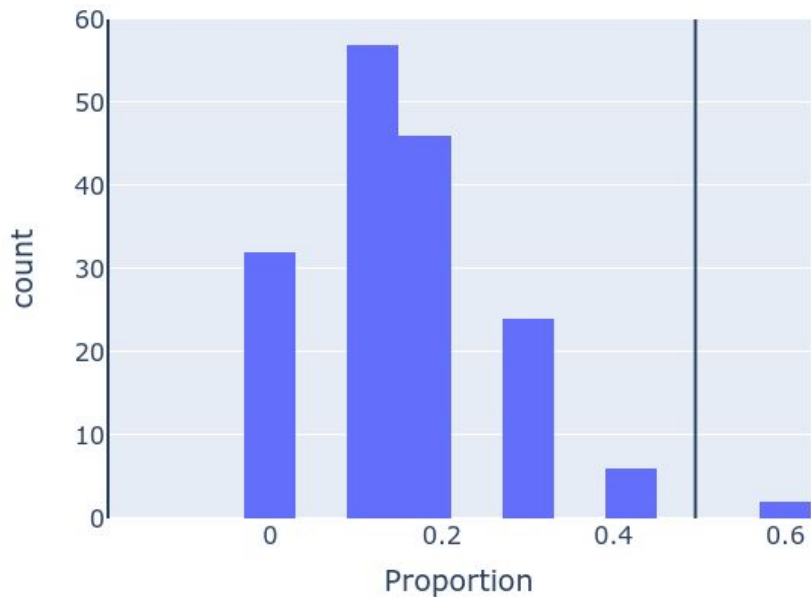


These predictions come from the XGBoost model. These predictions are done on 20% of the data that the model did not see during training. Only 15% of the predictions are wrong, and 85% of the predictions are correct. There's a slightly stronger tendency to predict INTP for any of the other personalities.

Predictions Over Time



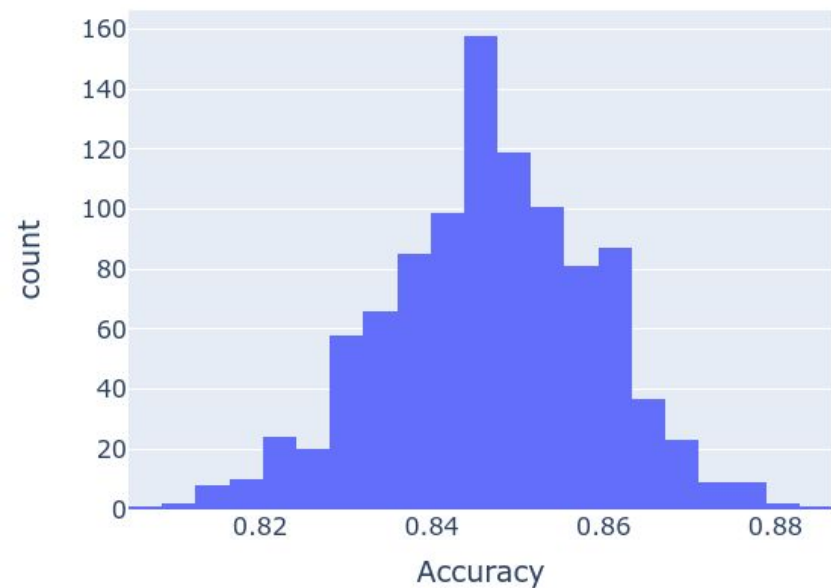
Histogram For Errors, 98.8% In Control



The errors are the fraction of 10 predictions that were wrong.

The errors are most likely to be 10%. Control limits were computed on the errors and we can see that the prediction error is mostly under control. There is a skew to the right which isn't good.

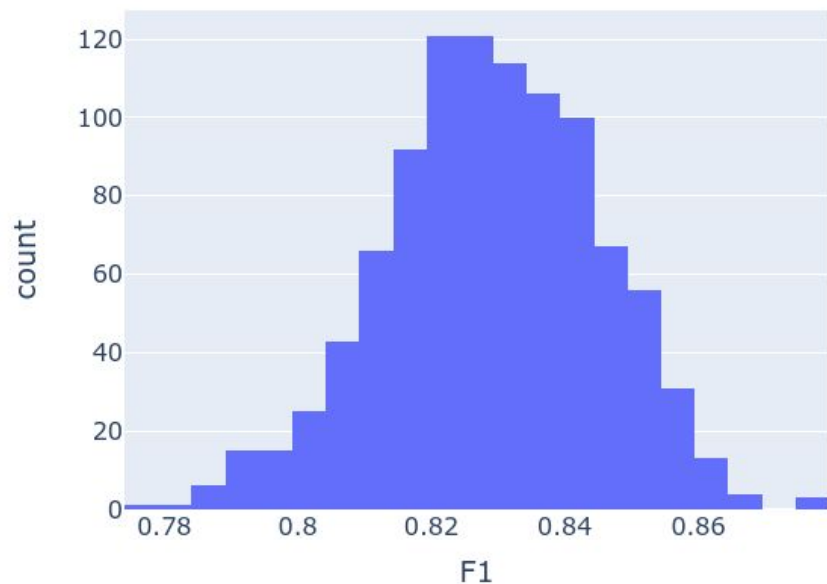
Histogram For Accuracy



The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then Accuracy was computed on each sample to get a distribution.

Accuracy has a tight range between 0.82 and 0.88, which is good.
Accuracy has a bell shape, which is good.

Histogram For F1



The prediction error was resampled as previously mentioned to get a distribution for F1. F1 is a combination of Precision and Recall. Precision tells us how well the model doesn't label another personality as a particular personality. Recall tells us how well the model finds all personalities.

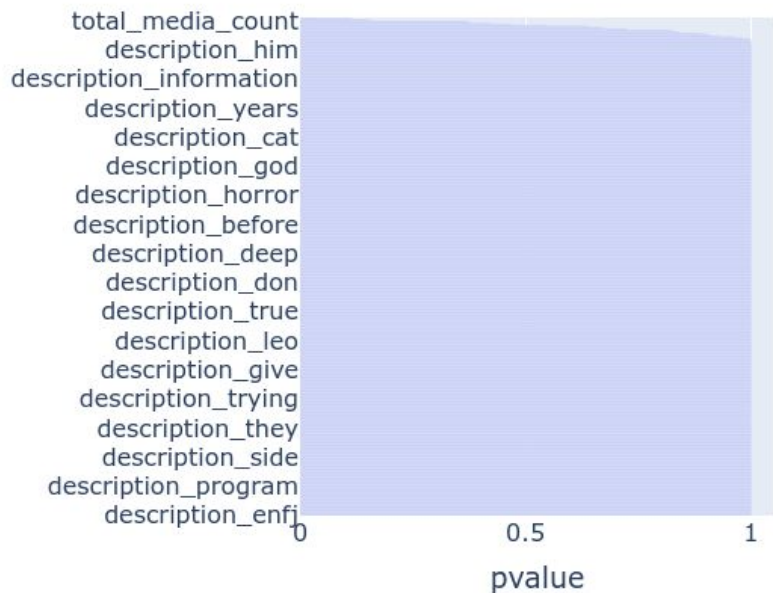
F1 has a reasonable range between 0.78 and 0.86, which is fine. F1 has a bell shape, which is good.

XGBoost Feature Importance



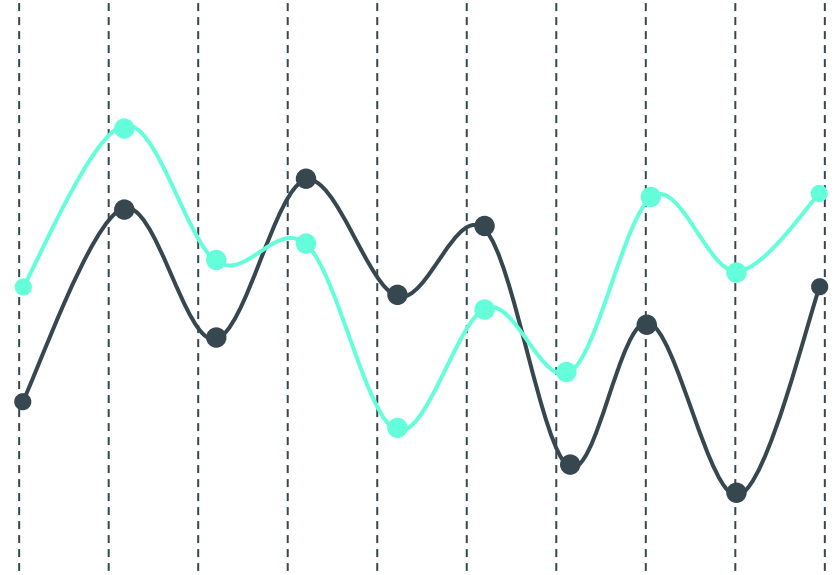
The top ten indicators of personality are shown to the left. We can see that the description may tell us the personality itself.

Feature Drift, Drift Detected If $pvalue < 0.05$



A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. The only column that experiences a drift is `average_media_count`. The model was retrained to include the test data.

Deployment



—

The data we start with.

The latest data we want
predictions for.

Retrain the model on
the initial data and new
data.



Data wrangling, feature
engineering, model
training.

See if the distribution of
the new data is
significantly different
than the initial data.

Thank You