Medium Popularity Prediction

Hsin

Outline

Dataset Introduction

Key EDA

KNN, Random Forest and XGboost

Final Model

Conclusion

Medium Dataset

1. **Source**: Kaggle, scraped from Medium
2. **Date:** Sep. 2017 to Sep. 2018
3. **Size of Data:** Converted from 279,577 to 100,000
4. **Reason for Extensive Data Cleaning**: Create a relatively small size data to test model more efficiently. Using random sample selection multiple times to ensure model accuracy.
5. **Goal of the project:** To help understand whether a post will be popular or not on Medium.

Dependant variable

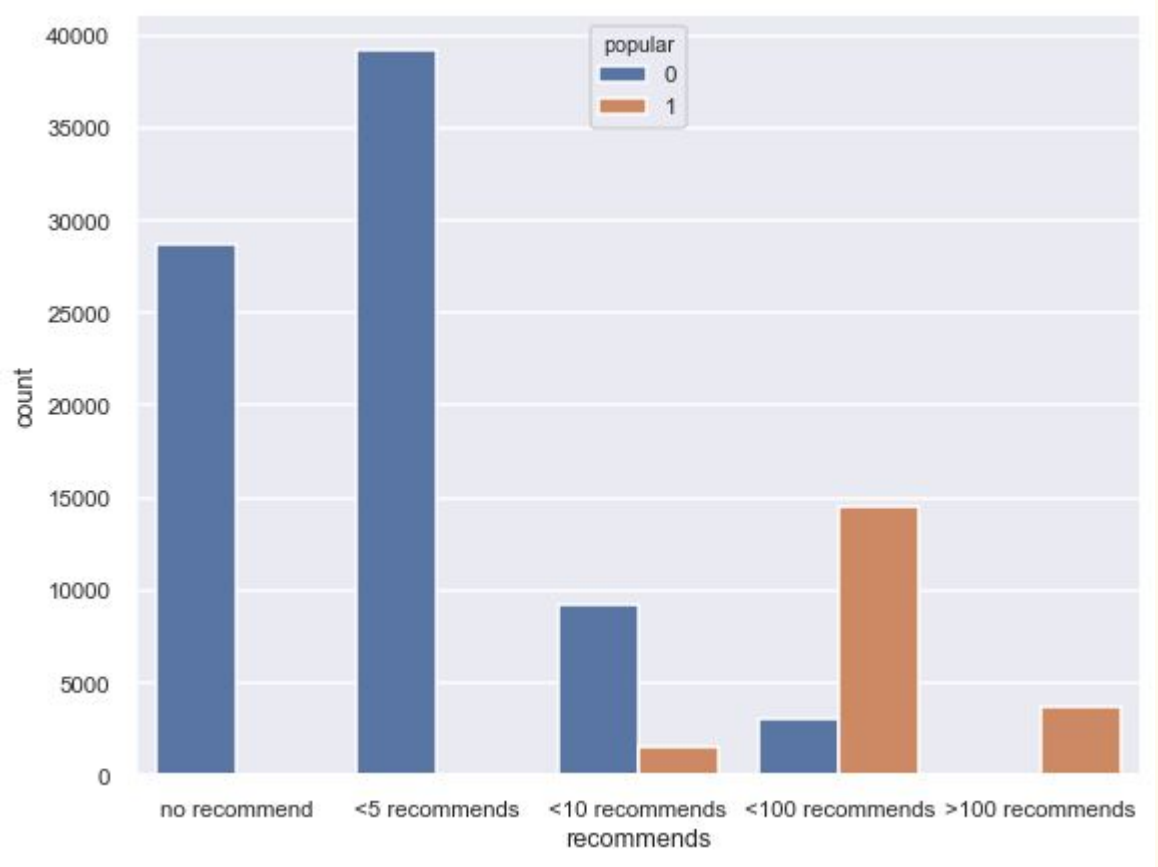
Popular:

Total number of claps and unique user who claps

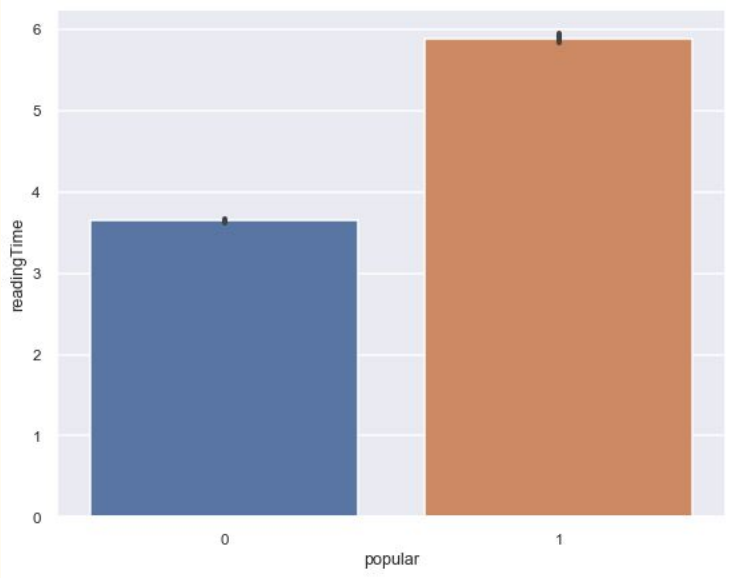
Bin together with 75% quantile

EDA

1. No popular article with recommends < 5
2. No disliked articles with recommends > 100
3. Recommendation has positive correlation to popularity

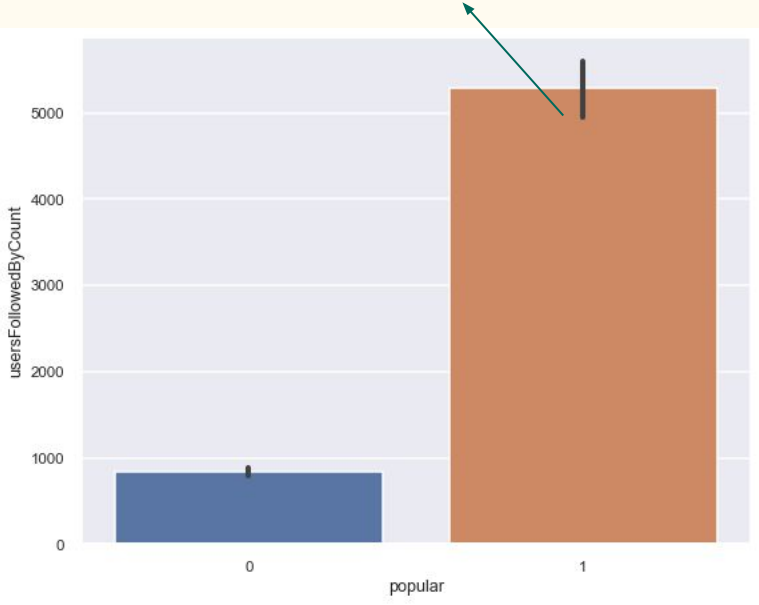


EDA



Popular posts tend to have longer reading time.

6.32 times more AVG followers in popular posts

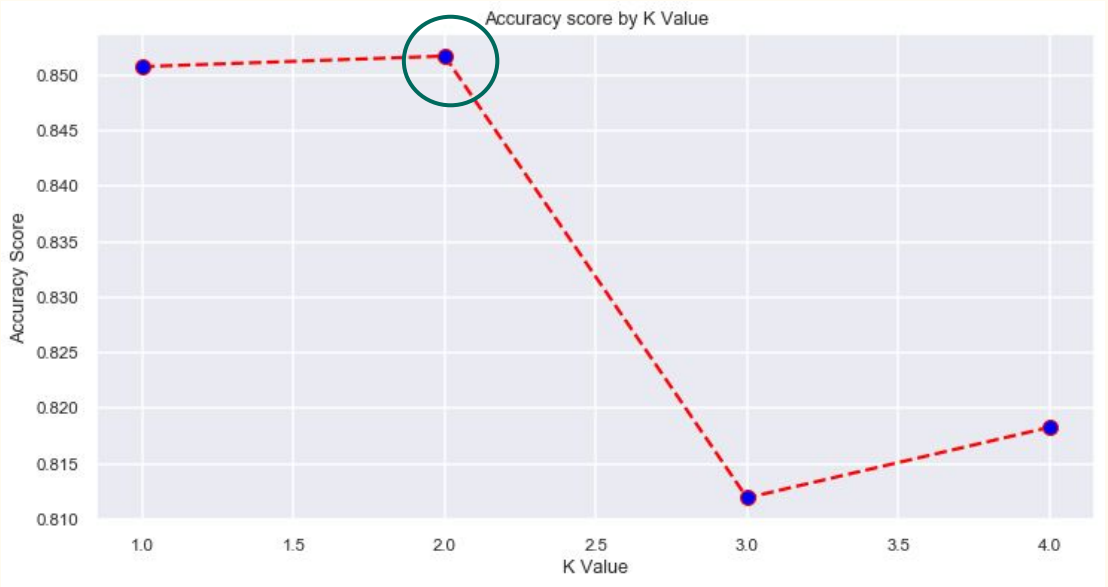


Followers significantly influence whether a post will be popular or not.

KNN

1. When K = 2, the accuracy rate is highest
2. Accuracy: 0.8517

   F1_score: 0.5984



Random Forest

Criterion = gini

Max depth = 10

Min samples split = 3

N estimators = 200

Accuracy: 0.8502

F1 Score: 0.6771

XGboost

N estimators = 1000

Learning rate = 0.5

Max depth = 10

Colsample bytree = 0.9

Min child weight = 2

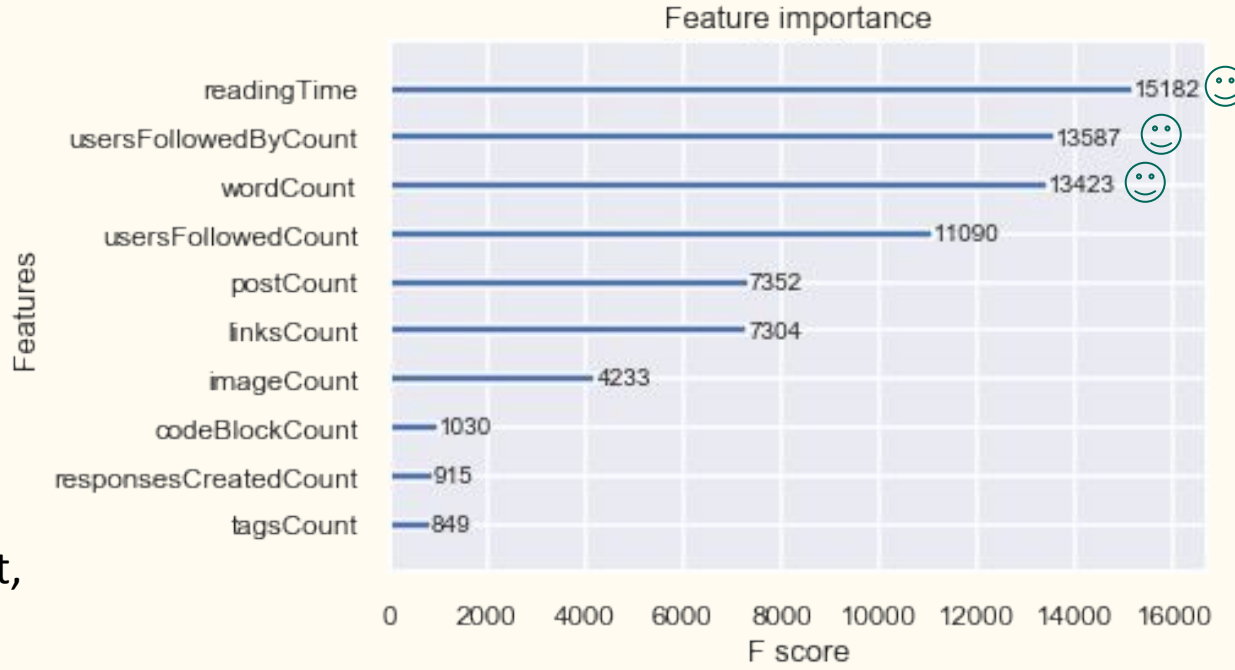
Accuracy: 0.9507

F1 Score: 0.8716

Final Model

XGboost is the final model

1. Top 3 important factors: reading time, followers, and word count impacts popularity most
2. Code amount, tag amount, and Number of responses to a post are least important features to popularity



Conclusion

1. In this project, a xgboost model is adopted to perform a high accuracy(95%) and both type 1 and type 2 error low(f1:87%) prediction.
2. By collecting informations from a given post, the model can tell the user whether it will be a popular article or not.
3. In conclude, reading time, followers, and word count impacts popularity most. For creating a successful blog post, users might consider to spend more time on these 3 parts.

Further Exploration

1. Followers might influence the model accuracy

2. Different tags included will be more useful

   (Data science v.s. Art)

Questions?