Project June

Andrew Seo

June Wine Bar – Brooklyn, NY: A Birthday

Introduction

- Motivation: Share the love
- How: Designing a classification model to pick “Good” wines
- Audience: Wine store employees and normal consumers

Tools & Methodology

- Data: Wine Enthusiast: Week of 6/15/17
- “Good” Wine >= Median Score
- Python: Analysis and Model Selection

Metric Decision

- Accuracy: Blindly classifying all as Good
- Recall: Knowing as many Good wines as possible
- Precision: Likelihood the wines recommended as Good are actually Good

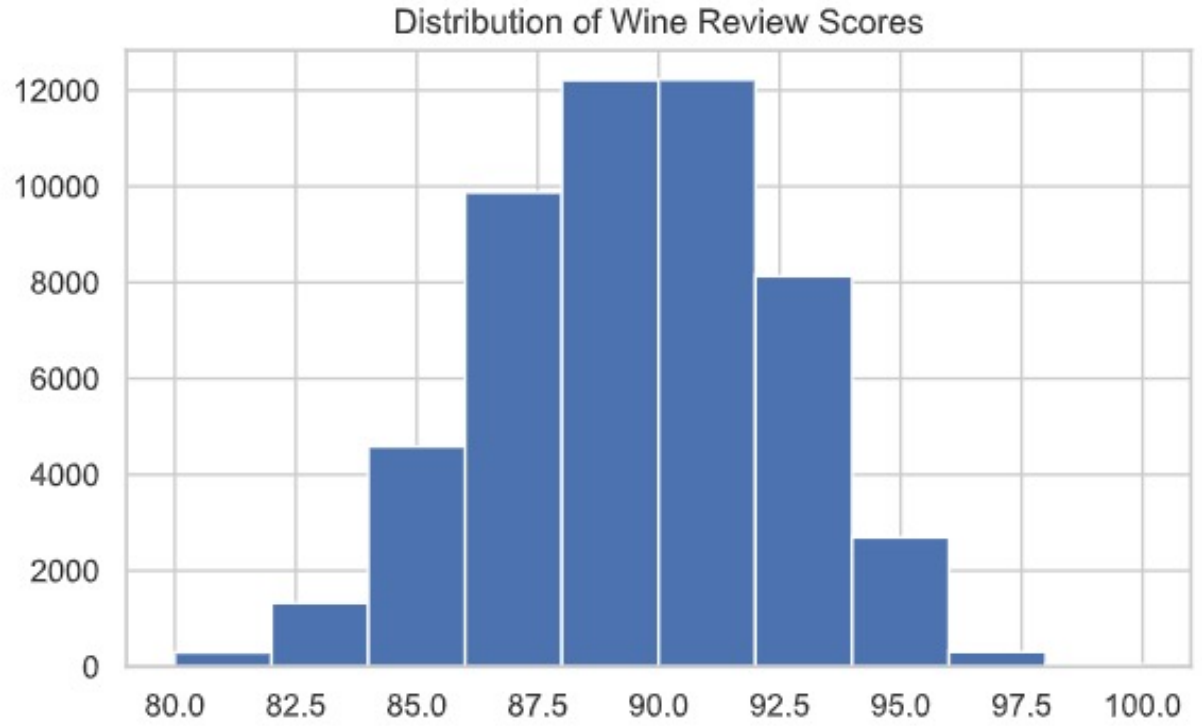
Goal: Precision

Why? Cost of false positive > cost of false negative

Picking a bad wine thinking it is good is worse than classifying an actually good wine as bad

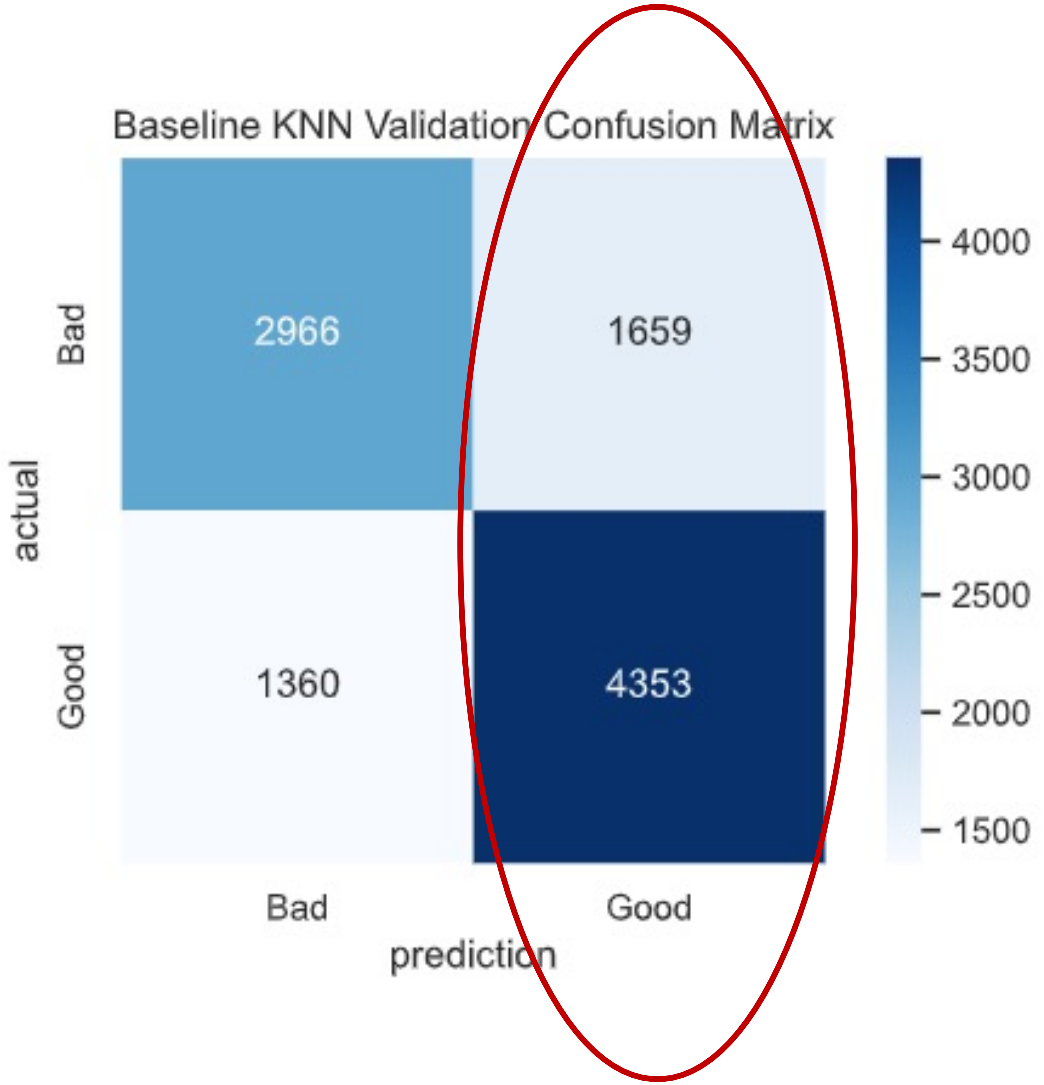
Data Cleaning

- Problem: Categorical Variables
- Solutions:
  - Grouping: “Other”
  - Dummy Variables: Names to Numbers



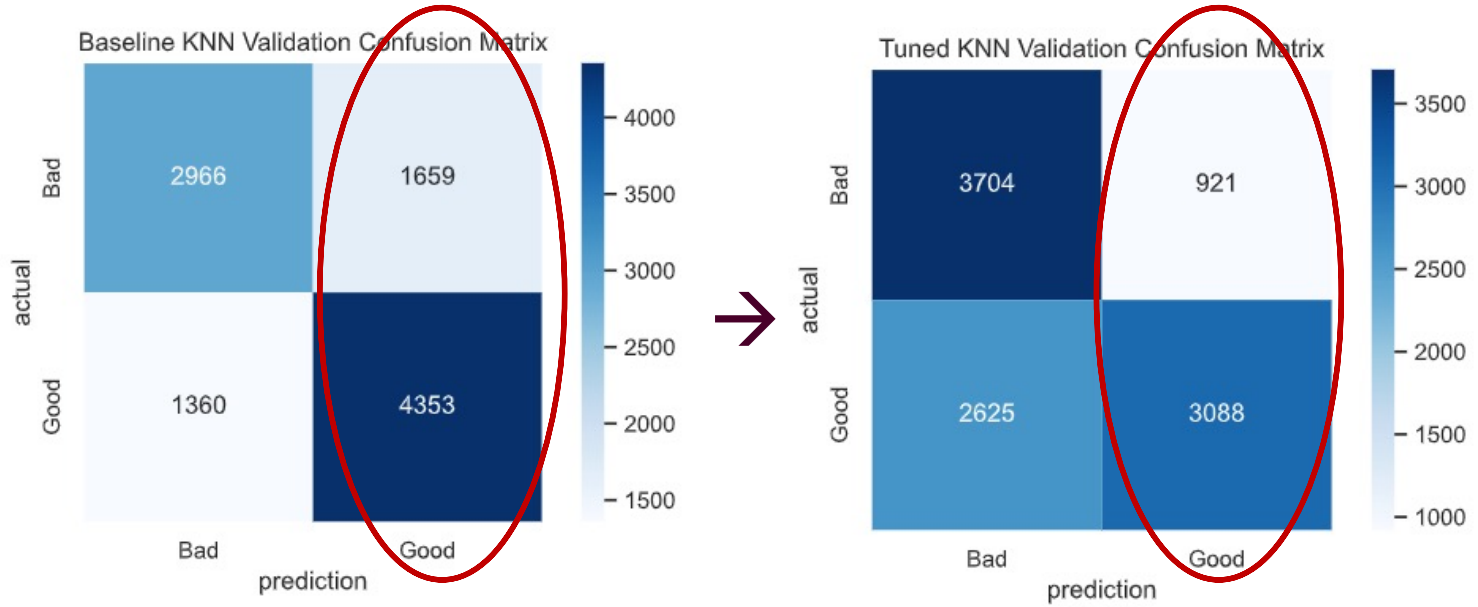
Analysis - Baseline

- KNN: 72.4%
- Logistic Regression: 75.2%
- Decision Tree: 74.0%
- Random Forest: 73.6%



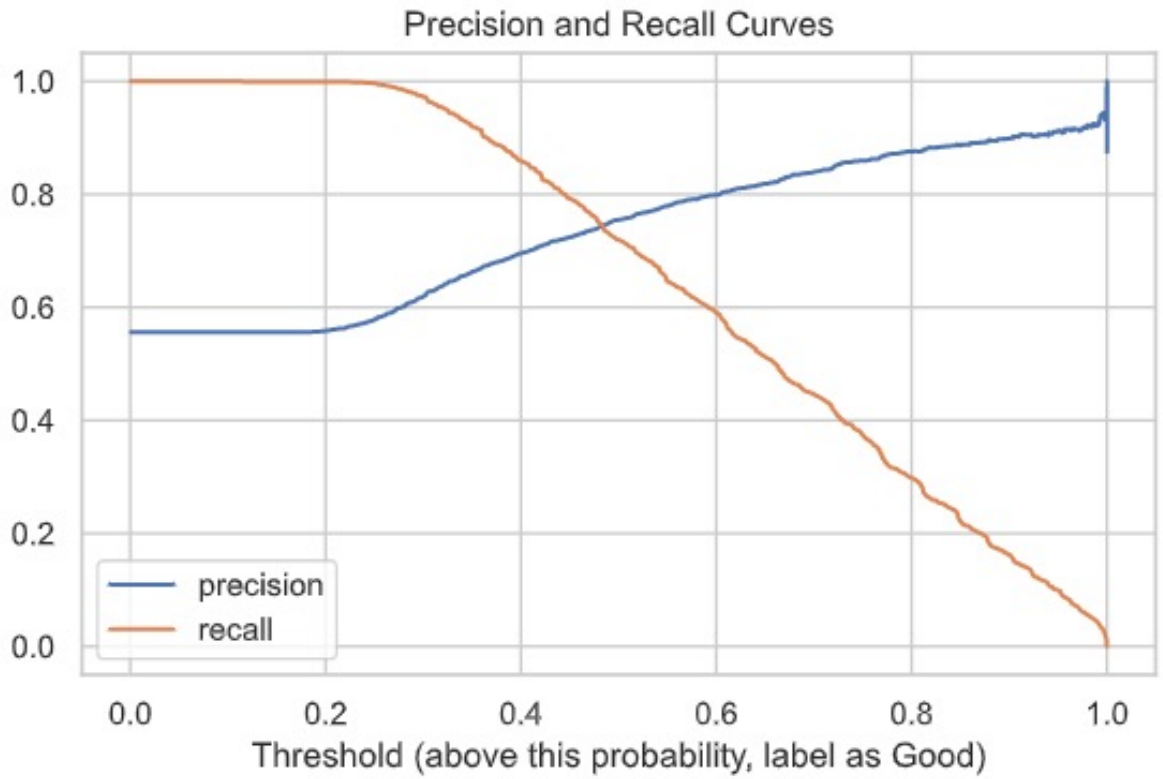
Analysis - Tuned

- KNN: 72.4% → 77.0%
- Logistic Regression: 75.2% → **84.0%!**
- Decision Tree: 74.0% → 75.9%
- Random Forest: 73.6% → 74.8%

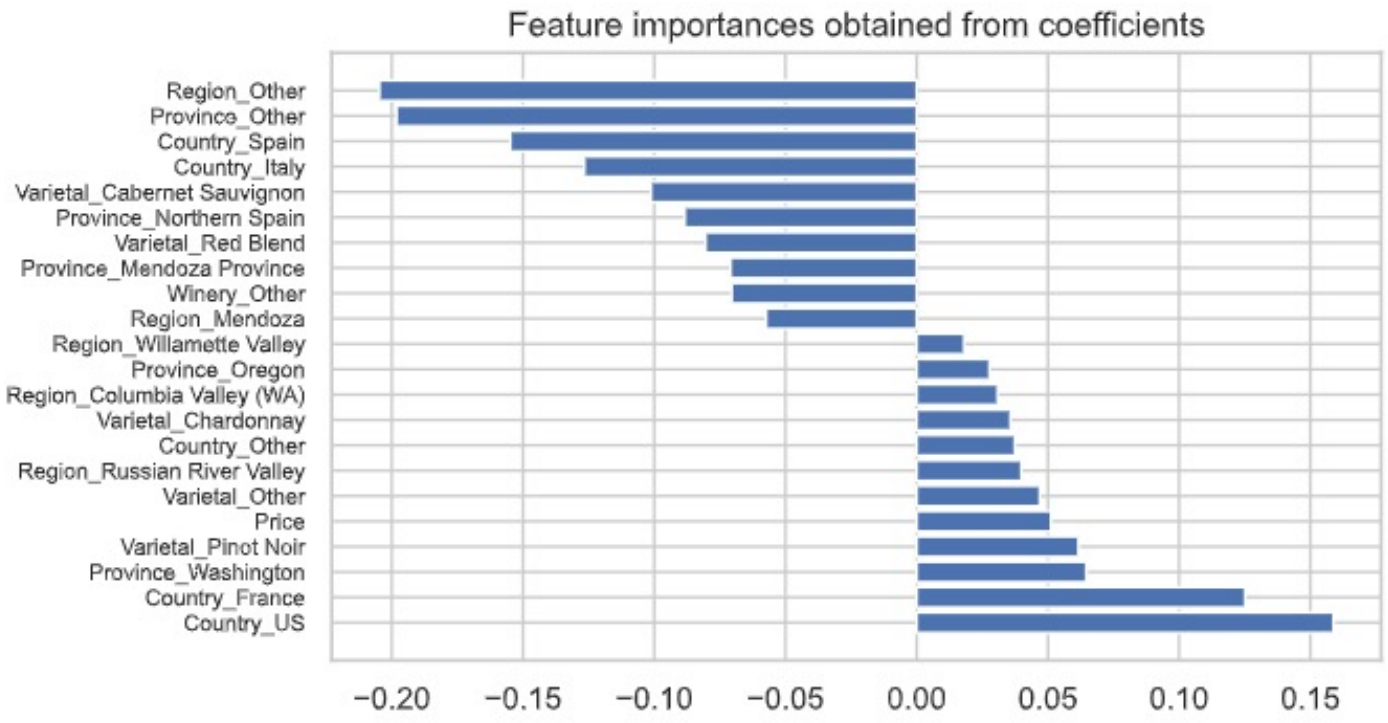


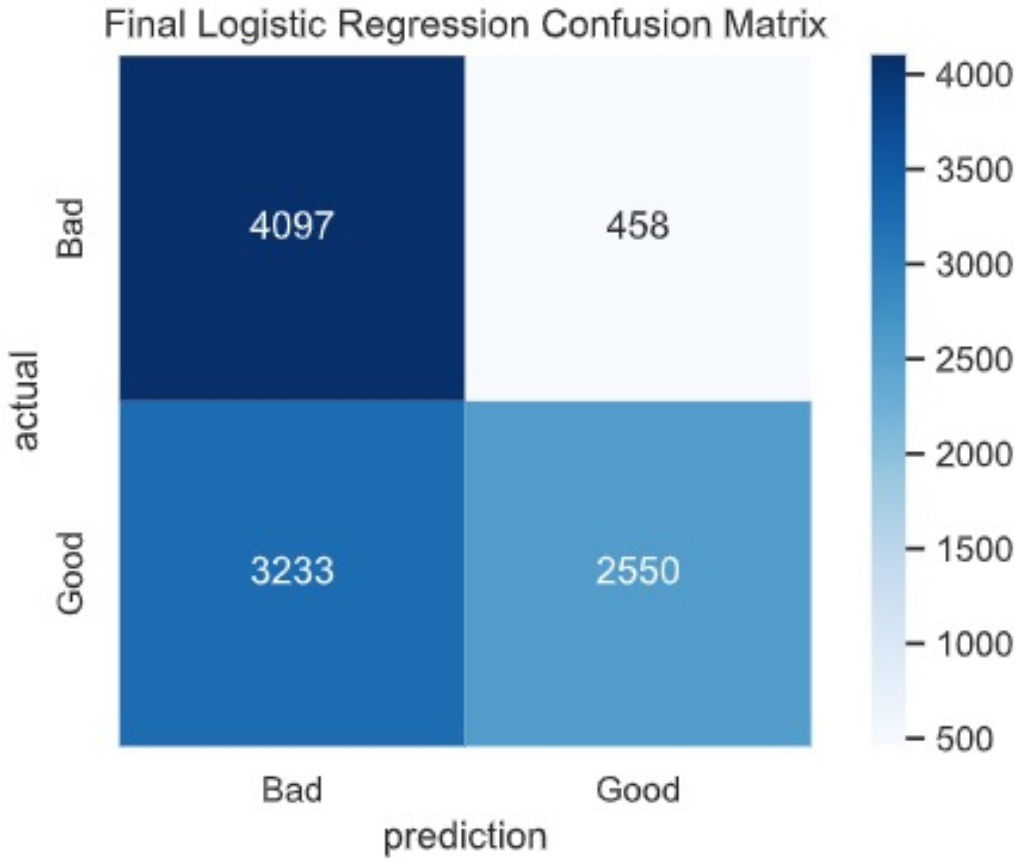
Analysis – Step Back

- Sanity Check: Other Metrics?
- Recall: 71.8% → 43.9%
- Of all of the actually good wines, the model identifies 43.9% of them (+12K wines)



Results





Precision: 84.8%

Conclusions

US, France > Italy, Spain

Pinot Noir > Cabernet Sauvignon

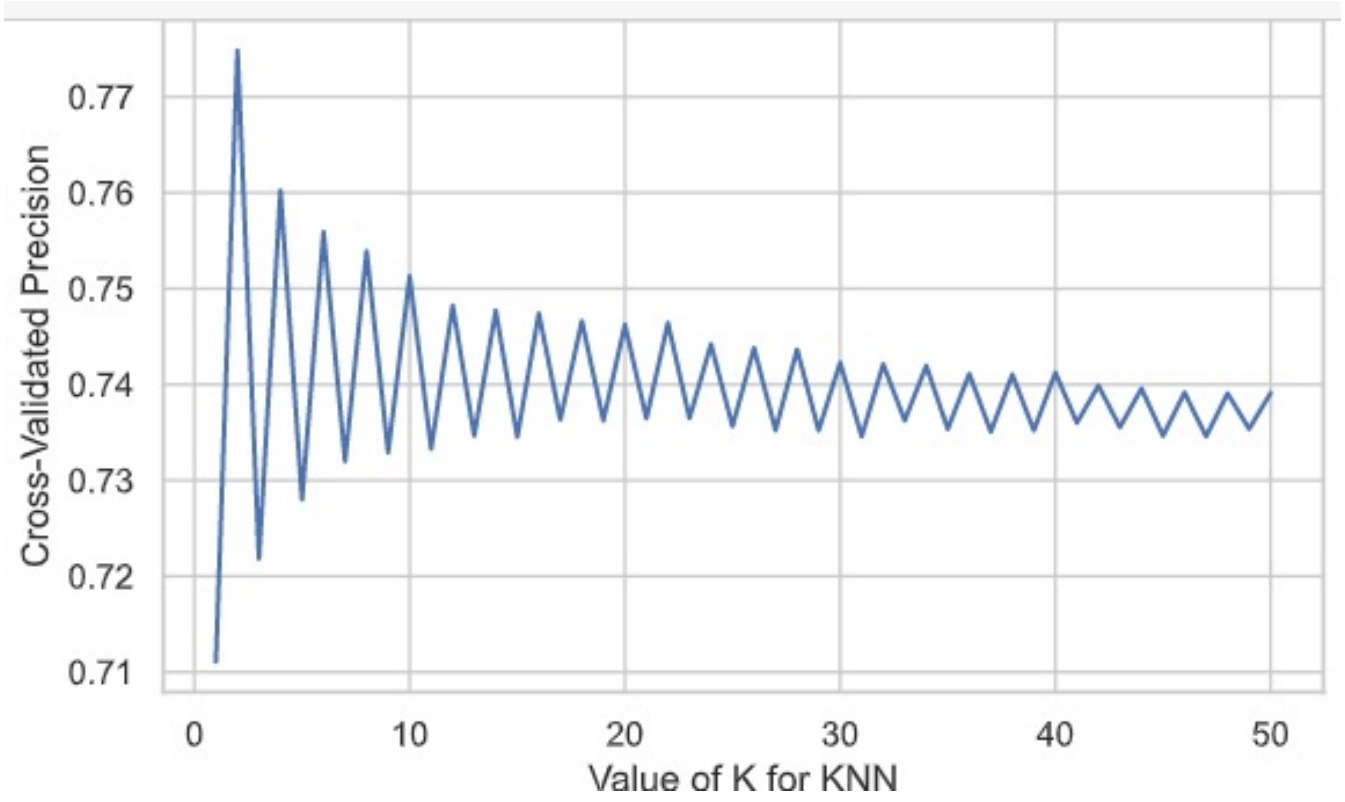
Price isn’t everything

Looking Ahead

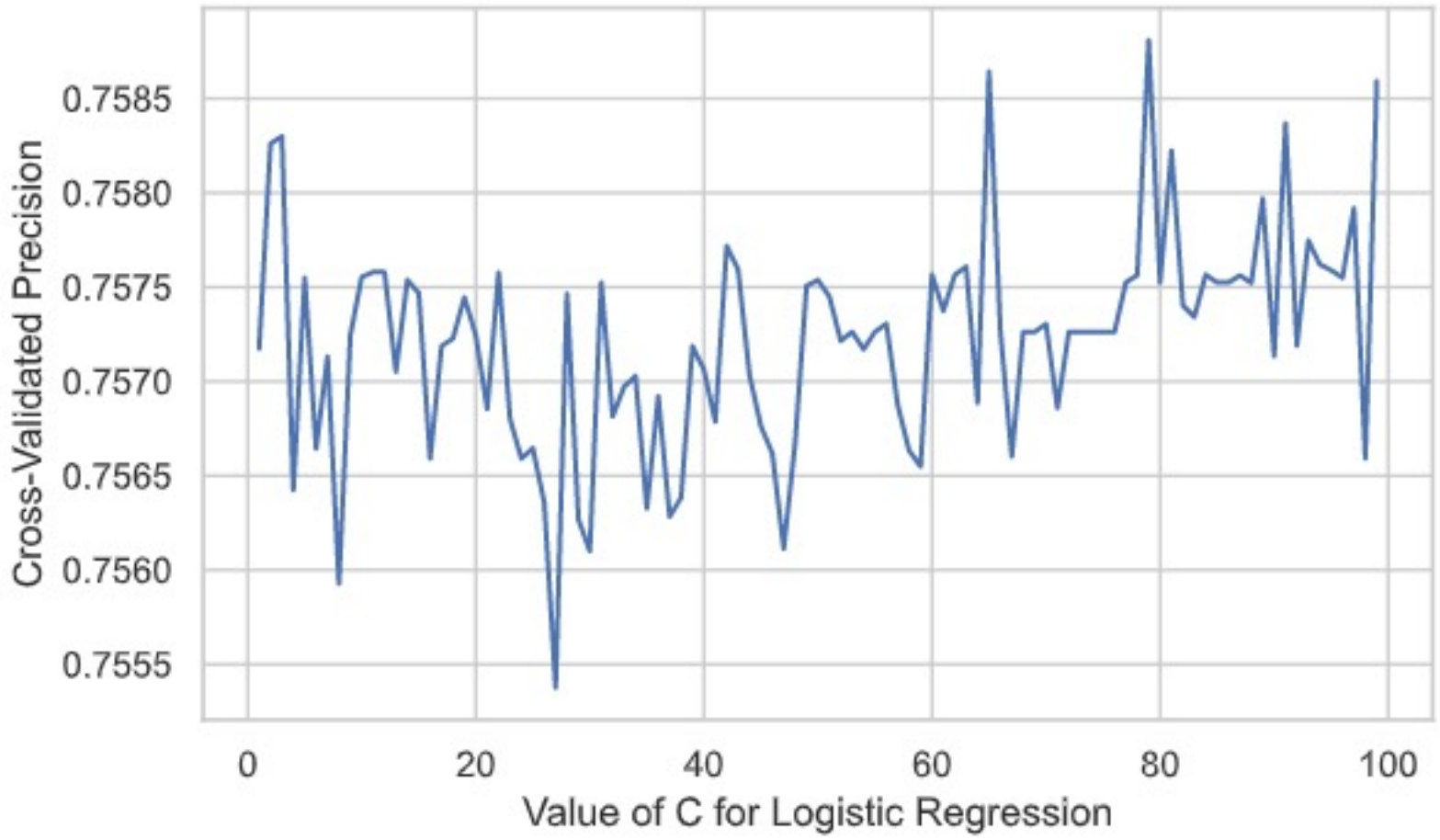
- Data: More quantitative/continuous variables such as pH, sulfate content, sodium level
- Models: Trying out XGBoost and Bayes models
- More comprehensive methodology: voting system

Questions?

Appendix


Cross-Validated Precision
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0
10
20
30
40
50
Value of K for KNN

Appendix: Cont.



Appendix: Cont.

Appendix: Cont.