Predicting Wine Quality: A Conundrum

Would you like some cheese with that?

Kalbi Zongo, Song Hoa Choi, Gina Shellhammer, Matt Edwards

June 2, 2014

Outline

Task

Predict the blind taster quality score of a wine based on chemical tests.

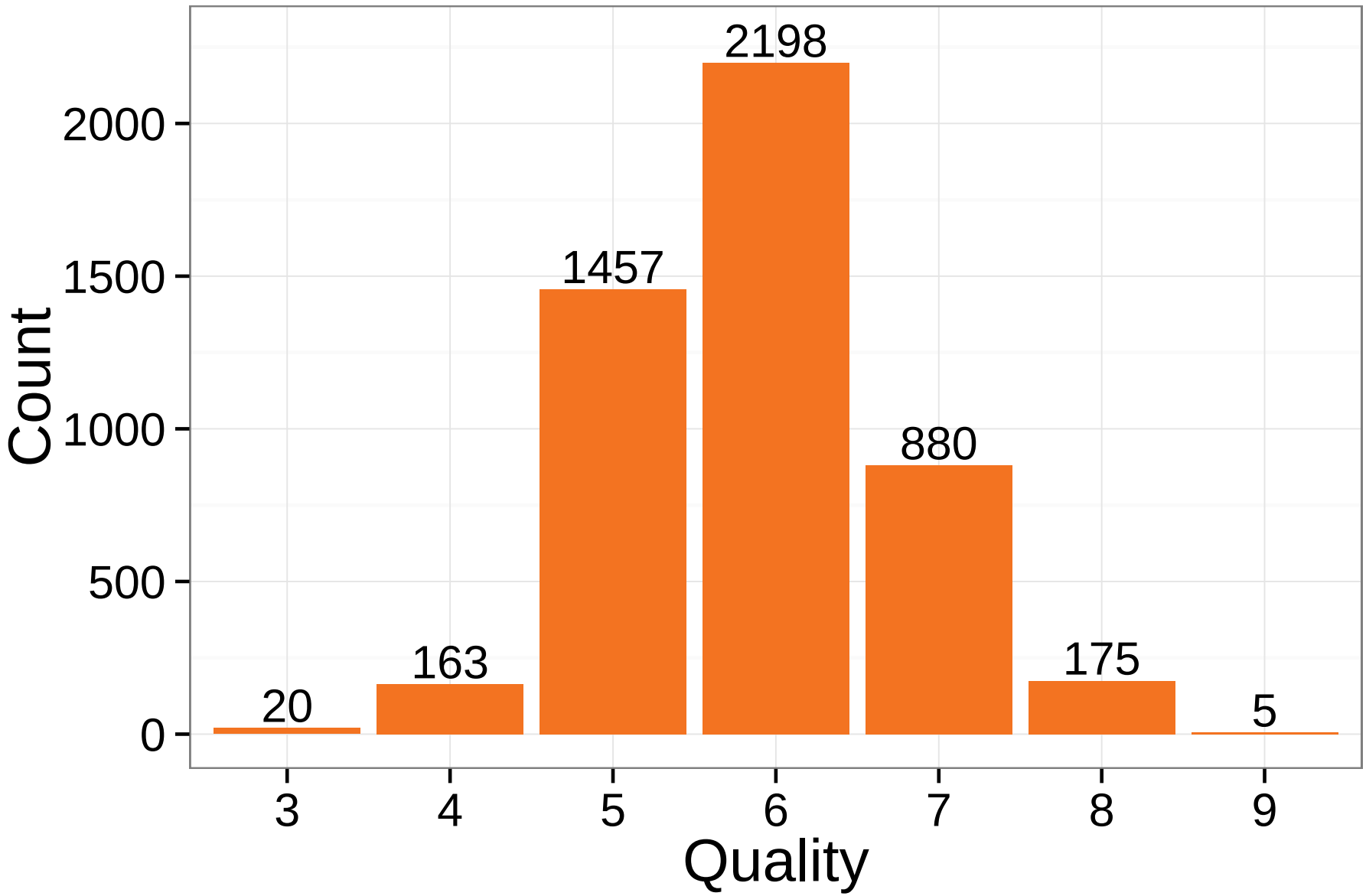
Data

- Two Datasets: Red & White vinho verde wine samples from northern Portugal
- 1599 & 4898 rows, respectively
- Concentrated on White Wine, due to more data

Data

- 11 Explanatory variables: measurements from various phytochemicals in wine
- Response variable "quality" is discrete variable on ordered scale from 0 (worst) to 10 (best)
- Nothing graded as 0, 1, 2, or 10

White Wine Quality Scores



Learning about Wine

Training and Testing Sets

- Training and Testing sets constructed through stratified sampling.
- Quality variable was the strata
- Why: Ensure representation of all quality categories in both Training & Testing datasets.
- How: 37.5% of items (rounded up) in strata were randomly selected to be in the testing set. Remaining 62.5% were the training set.

Regression

- With "prediction" as the goal, we think regression.
- Forward, Backward and Subset Model Selection done, all resulted in same model.
-

K-Nearest Neighbor Regression

- using some measure of distance, find nearest neighbors in dataset
- Order examples by increasing distance
- Find a "optimal" number k of nearest neighbors
- Calculate an inverse distance weighted average with the k-nearest multivariate neighbors
- Used `rminer` package in R. Offers many regression types

Ordinal Regression

Also "Ordered Logistic" or "Probit" Regression

- using some measure of distance, find nearest neighbors in dataset
- Order examples by increasing distance
- Find a "optimal" number k of nearest neighbors
- Calculate an inverse distance weighted average with the k-nearest multivariate neighbors

Classification

Logistic regression used as classifcation to predict category
Split data into 3 sets: Training, Cross-Validation, Test Set - completely new.
polynomial logisitc regression, fit 4 degree poly nomial model, fit a shrink parameter tried 12 to avoid overfitting. estimated the training set error, picked the best and tried it on the cross validation set. then used that parameter to predict with the test set.

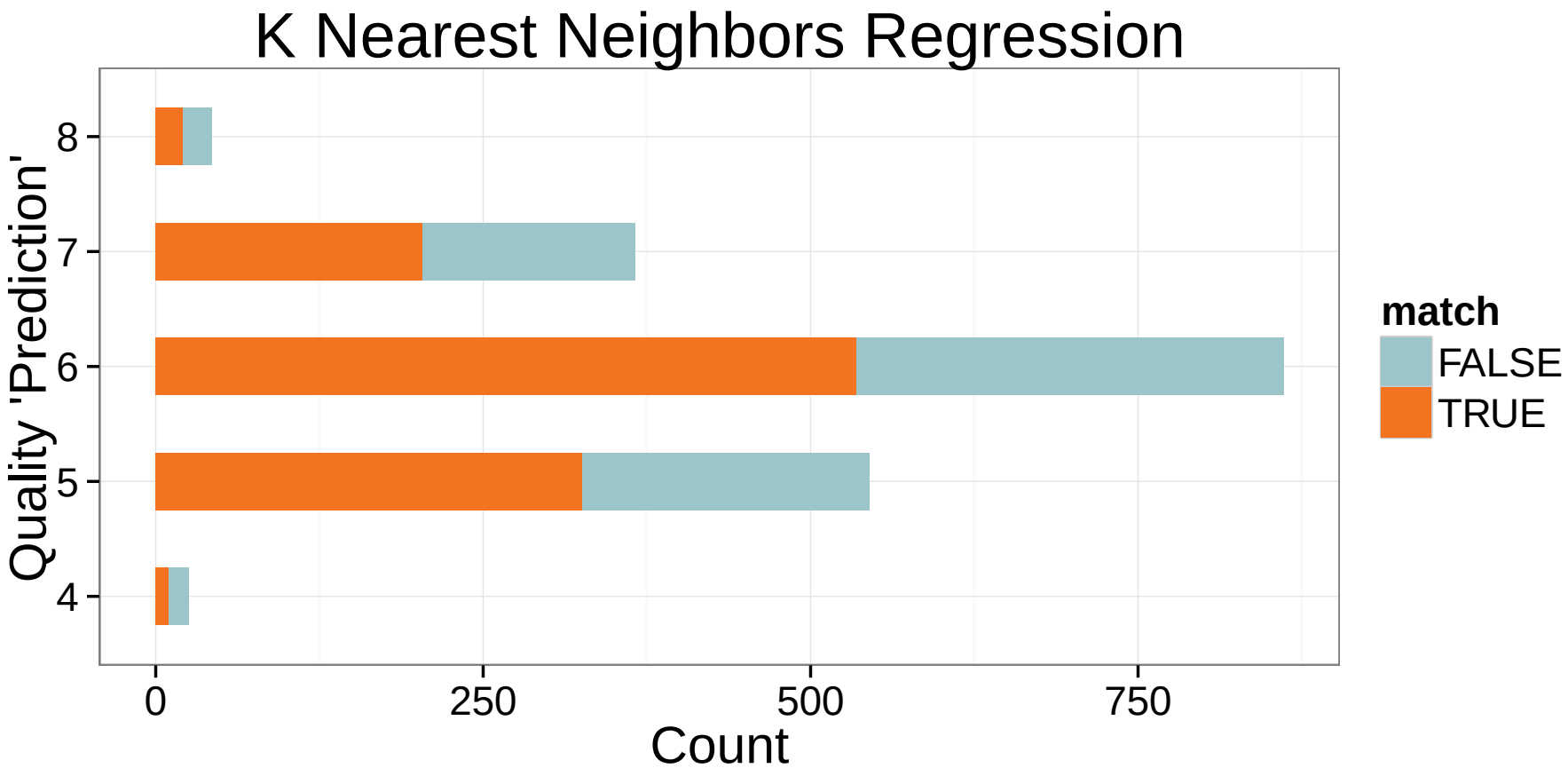
Random Randomness is Random

- 75% of Quality ratings were either 5 or 6.
- Is randomly assigning 5 or 6 to everything as good as, or better than, our other methods?
- Using `rbinom(1,1,0.6014)`, 1s were predicted as quality 6, 0s as quality 5
- Probability of 60.14% because from Training Set, considering only 5s and 6s, 6s were 60.14% of total observations
- Our base line success rate to compare other methods.

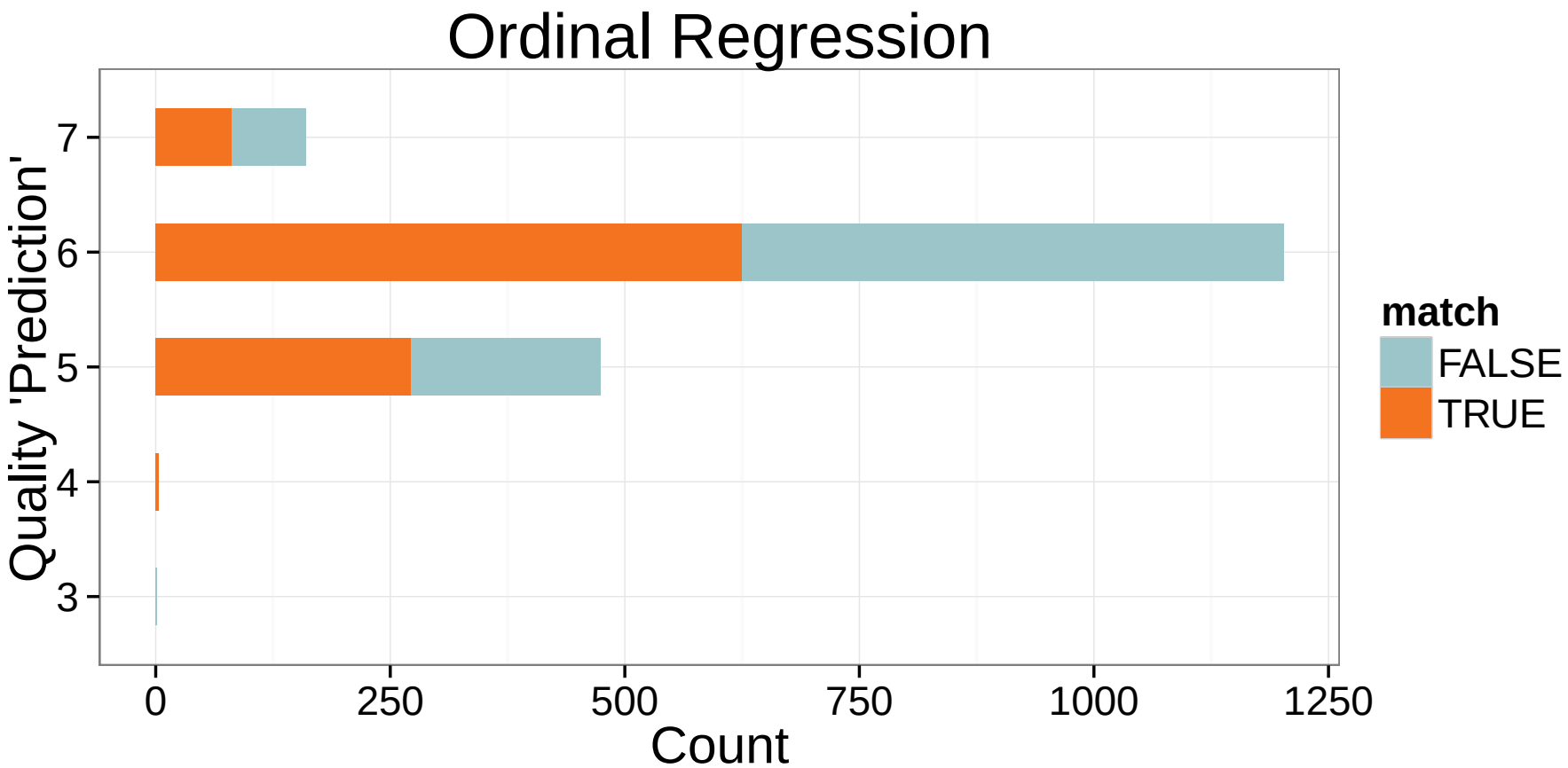
Results


pale yellow core
rich gold core
primary colors
primary colors

K Nearest Neighbors Regression: 59.6% Success Rate



Ordinal Regression: 53.3% Success Rate



Regression Summary

- K Nearest Neighbors
  - Overall 59.6% success rate.
  - No properly allocated 3s or 9s
- Ordinal Regression
  - Overall 53.3% success rate
  - No properly allocated 3s, 8s or 9s.

Regression: Success by Quality Predicted

	KNN		Ord		
Prediction	% Match	% Fail	% Match	% Fail	# Present
3	n/a	n/a	0%	100%	8
4	40.0%	60.0%	100%	0%	62
5	59.8%	40.2%	57.4%	42.6%	547
6	62.1%	37.9%	52.0%	48.0%	825
7	55.7%	44.3%	50.6%	49.4%	330
8	48.8%	51.2%	n/a	n/a	66
9	n/a	n/a	n/a	n/a	2

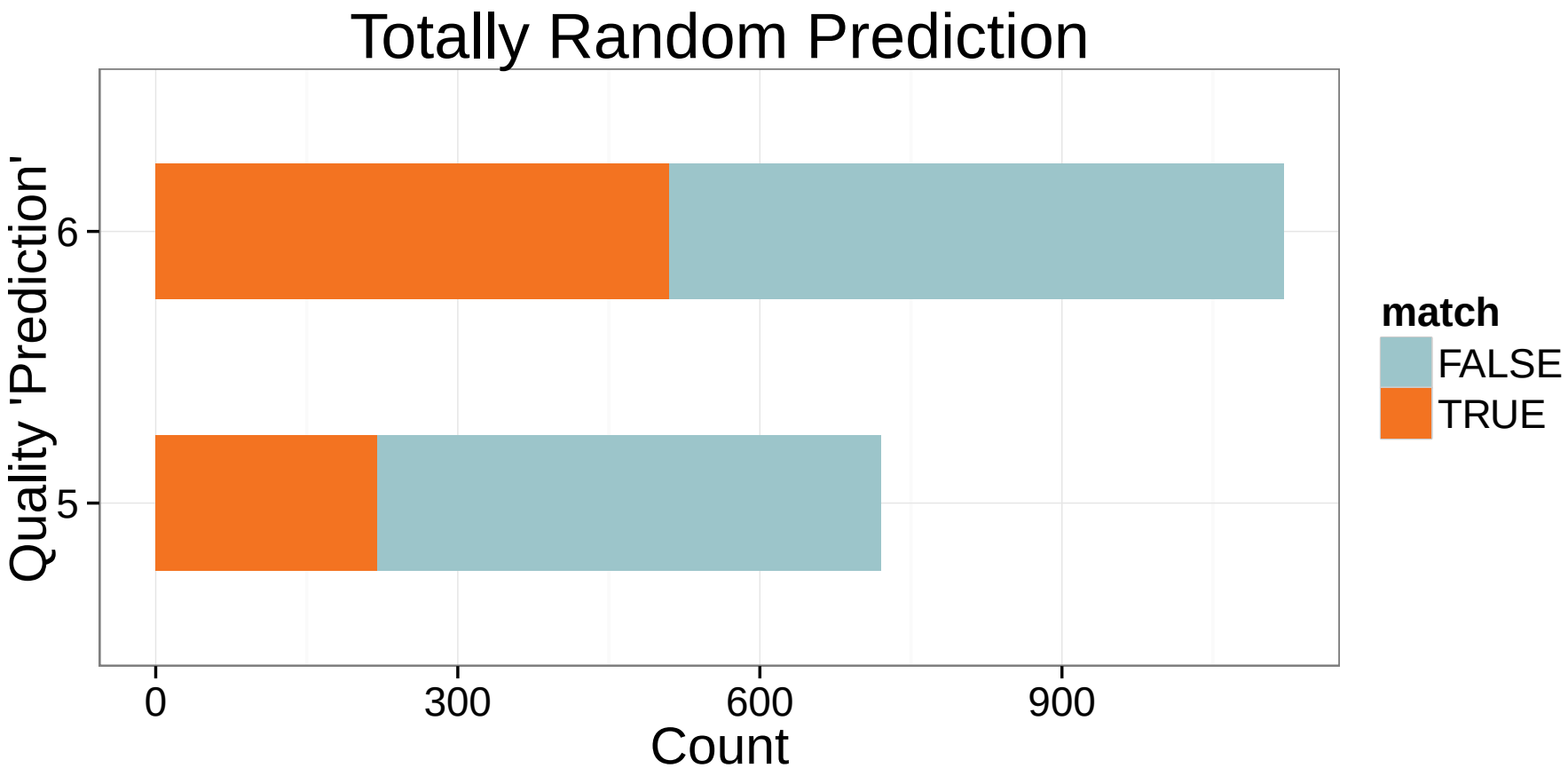
Success by predicted quality. 'N/A' means nothing predicted at that quality.

Classification: 50% Success Rate

content... change success rate in title

Random 'Prediction': 39.67% Success Rate

Turns out, that's not really a great 'prediction' method. Who knew?



Discussion


ELKI
LAS MORAS
GCHA-MAM
TORRONTES
LE CHANAY
Hubert Lamy
BOURGOGNE
2009
CÔTES DE GASCOG
CRÉMANT DU JURA
CHARDONNAY
2009
MORA

Cross Validation
Search method for `knn`
Ordinal Regression - must assume that ordered categories - assumption about ordering of categories has some repercussions.
Ordinal Regression: proportional odds - coefficients stay the same, and the intercept value changes. all explanatory variables have the same weight for all categories. Puts them in possible categories, picks the one with the highest probability.
knn - if category distribution is skewed, larger categories can dominate, which is what we see in our results.

Questions