Wine quality classification using logistic regression: A comparison of ordinal and non-ordinal methods

Tsering Yangchen lama
School of Computing and Information Technology
University of Wollongong
CSCI933, SN:7595499, tyl936@uowmail.edu.au

April 24, 2023

Abstract

This report detailed study of building regression models on quality of wines. Regression model is to predict dependent variable based on multiple independent variables. Logistic regression had been chosen the subject to study based on the provided dataset which is more on classification. Furthermore, scikit learn library build a multinominal logistic regression whereas stats model library has probit and logit ordinal regression. Moreover, both had been compared to find the best test accuracy for the performance of experiment.

1 Introduction

Wine quality is important factors for consumers' choice and purchasing decision. comprehension towards wine quality and ability towards prediction of wine quality will give valuable information for consumer, retailers, and producers. In this report, logistic regression is experimented along with regularization technique such as L1, L2 regularization and elastic net regularization. It is implemented with help of python libraries such as scikit-lean (sklearn) and statsModels. For Sklearn, wine quality is further categorised into low, medium, and high. As Sklearn is more toward classification and multinominal. Similarly, wine quality for stats model as in ordinal nature which is wine's rating include 1-10 . Lastly, these models are compared and predicted for the best accuracy. Overall, this report aims to provide a comprehensive comparative analysis of different binary classification methods, including logistic regression with regularization, probit, and logit models, for predicting wine quality ratings. The findings of this study can offer valuable insights to wine industry professionals, researchers, and wine enthusiasts, and contribute to a deeper understanding of the factors that influence wine quality ratings

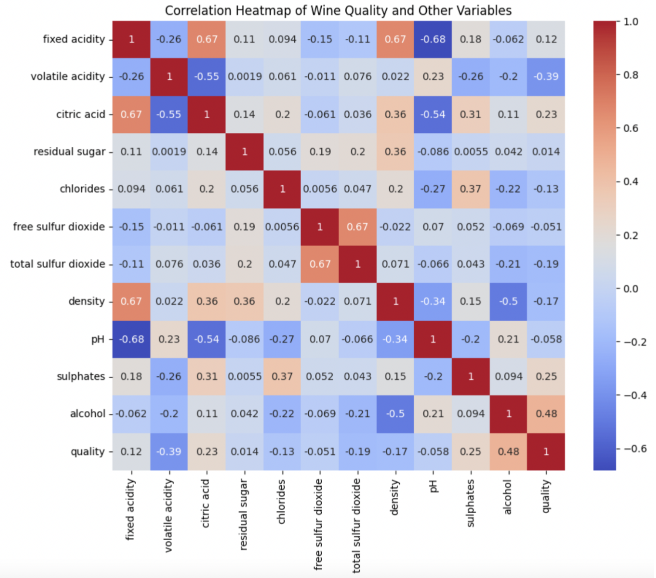
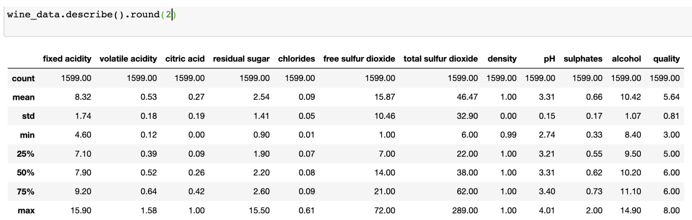


Figure 1: Description of wine quality co-relation

2 Wine quality dataset

According to the wine quality dataset prepared by Cortez, Cerdeira, Almeida, Matos, and Reis (2009), there are 11 columns which represents different attributes of a wine, including fixed acidity, volatile acidity, citric acid, residula sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alocohol. There are total 1599 rows which represent individual samples of wine. Good thing is that it doesnot have any missing values as seen from non-null count.For quality of wine, it has rating from 0-10 but only ratings of 3, 4, 5, 6, 7, 8 have found in this dataset. Based on the correlation values in figure 1, Alcohol and volatile acidity have higher absolute correlation values with wine quality. It means that higher the alcohol values, better the quality is. The least co-related values is residual sugar which is only of 0.013.

2.1 The descriptive statistics report of wine Data



```
wine_data.describe().round(2)
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.00	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	22.00	1.00	3.21	0.55	9.50	5.00
50%	7.90	0.52	0.26	2.20	0.08	14.00	38.00	1.00	3.31	0.62	10.20	6.00
75%	9.20	0.64	0.42	2.60	0.09	21.00	62.00	1.00	3.40	0.73	11.10	6.00
max	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00

Figure 2: Description of wine dataset

- The mean fixed acidity of the wines in the dataset is 8.32, with a standard deviation of 1.74. This indicates that the majority of wines in the dataset have a fixed acidity around 8.32, but there is some variability in this feature among the wines.
- The std row shows the standard deviation of each column, which provides a measure of how spread out the data is around the mean. For instance, the ”alcohol” feature of the wine dataset, the standard deviation of this feature can give us an idea of how much the alcohol content varies across different wines. A high standard deviation would indicate that the alcohol content varies significantly among the wines, while a low standard deviation would indicate that the alcohol content is relatively consistent among the wines.

3 Theory and properties of regression

Regression is the subcategory of supervised learning which is a method for modelling and analysing the interactions among variables and how they play a role to producing a specific response together. The concept of fitting a mathematical model to observed data to make predictions or understand the underlying relationships between variables supports regression theory. The most popular method of regression is linear regression, which implies that the dependent variable and the independent variables have a linear relationship. When the relationship between variables is more complex, other types of regression, such as polynomial regression and logistic regression, can be used as per Pedregosa et al. (2011). For multiple linear regression, which has more than one independent variable, the formula becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

where:

- Y is the dependent variable or response variable
- $X_1, X_2, \ldots, X_k$ are the independent variables or predictor variables
- β_0 is the intercept or constant term
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the independent variables
- ϵ is the error term, which represents the unexplained variation in Y that is not accounted for by the model Chen (2019).

The properties of regression include:
Linearity: It is assumed that the relationship among the dependent variable and the independent variables is linear. The observations are independent of each other. **Homoscedasticity**: The error variance is constant across all levels of predictor variables. There is no multicollinearity because the predictor variables are assumed to be uncorrelated. The independent variables' effects on the dependent variable are assumed to be additive Chen (2019).

3.1 L1, L2 and Elastic-net penalties

Concerned regarding overfitting which are raise due to unfamiliarity of input features to the output are well handled with the help of following penalties.

The Lasso regularization, which is also called the L1 penalty, involves introducing a penalty component into the loss function that is proportional to the absolute value of the coefficients of the model. This causes the coefficients to be compressed towards zero, making it useful for feature selection by setting some of the coefficients to zero.

The L2 penalty, also known as Ridge regularisation, introduces a penalty term into the loss function that is proportional to the model coefficients squared. This has the same effect as L1 regularisation in that it shrinks the coefficients towards zero, but in a gentler way Kim, Koh, Lustig, Boyd, and Gorinevsky (2007).

the Elastic-net penalty is the combination of L1 and L2 penalties, provides a way to balance their effects. It adds a penalty component to the loss function that is a linear combination of the L1 and L2 penalties, with the weight of each penalty adjustable via a hyperparameter. This method is especially useful when certain features have strong correlations, because the L1 penalty tends to select only one of the correlated features, whereas the L2 penalty assigns comparable coefficients to both Friedman, Hastie, and Tibshirani (2010).

3.2 Logistic regression

Logistic regression, also referred to as logit regression, maximum-entropy classification (MaxEnt), or log-linear classifier, models the probabilities of different outcomes of a single trial by utilizing a logistic function. This implementation has the capability to accommodate binary, One-vs-Rest, or multinomial logistic regression, and can include optional regularization such as Ridge, Lasso, or Elastic-Net. The binary case of logistic regression can be extended to multiple classes, resulting in the multinomial logistic regression.Simon, Friedman, and Hastie (2011)

Logistic regression is a significant proportion of GLMs that utilizes a Binomial/Bernoulli distribution and a Logit link function. The model produces a numerical output that represents the predicted probability, which can then be used as a classifier by applying a threshold (typically 0.5) Minka (2003). Logistic Regression is used as a classifier in scikit-learn because it requires a categorical target.

3.3 Ordinal logistic regression

As per IBM (2021), Ordinal regression is a statistical method that enables to analyze how an ordinal response variable (a variable that has an ordered set of categories) depends on a set of predictors, which can be either categorical or continuous variables. The method is based on the work of McCullagh, and in the software syntax, it is referred to as PLUM. Essentially, it allows to model the relationship between the ordinal response variable and the predictors, and estimate the probabilities of each possible outcome category given the values of the predictors.

3.4 probit an logit ordinal regression

Logistic and probit regression are both examples of generalized linear models. They share the same mathematical form and can be used to model the relationship between one or more predictor variables (numerical or categorical) and a categorical outcome variable. Both types of models have versions for binary, ordinal, or multinomial outcomes, and require specific coding of the outcome variable Grace-Martin (2018).

The main difference between the two models is their theoretical basis. Instead of using the outcome variable directly, generalized linear models use a function of the mean of the outcome variable, which is called the link function. In logistic regression, the link function is the logit, which is a type of logarithmic transformation.

$$g(p) = \log\left[\frac{p}{1-p}\right]$$

where $g(p)$ is the logit function of the probability p, and p is the probability of an event occurring (i.e., the outcome variable).

In contrast, probit regression uses an inverse normal link function.

$$g(p) = \Phi^{-1}(p)$$

where $g(p)$ is the probit function of the probability p, Φ^{-1} is the inverse of the standard normal cumulative distribution function, and p is the probability of an event occurring (i.e., the outcome variable).

4 Experiments

Five random data from wine quality dataset is set aside, which will later use for testing. Remaining dataset is leveraged for further exploration.For sklearn model, Conversion of ordinal wine's quality into multi-nominal wine's quality is done by adding a new column called `quality_cat` and then grouping the `"quality"` column into four bins and assigning them labels `"A"`, `"B"`, `"C"`, and `"D"`.Category of wine has four classification.

For statsModels, quality of wine is in ordinal order including **low**, **medium** and **high**. Next, scaling the input variables using the StandardScaler function from `scikitlearn`. By standardizing the features, it ensurer that dataset have the same scale and range, allowing the model to weigh them equally. This function standardizes the data so that it has a mean of 0 and a standard deviation of 1.

4.1 Data split

`train_test_split` from sklearn library assists to divide the wine data into Training and testing set. For this experiment, there are 1275 data for training and 319 for testing purposes. The test size is set to 0.2, which means that 20% of the data is used for testing and the remaining 80% is used for training. The random state is set to 100 for reproducibility. `ordinal_train` and `targetordinal_train` contain the training set of the ordinal features and target variable respectively, while `ordinal_test` and `targetordinal_test` contain the test set of the same features and target variable respectively.

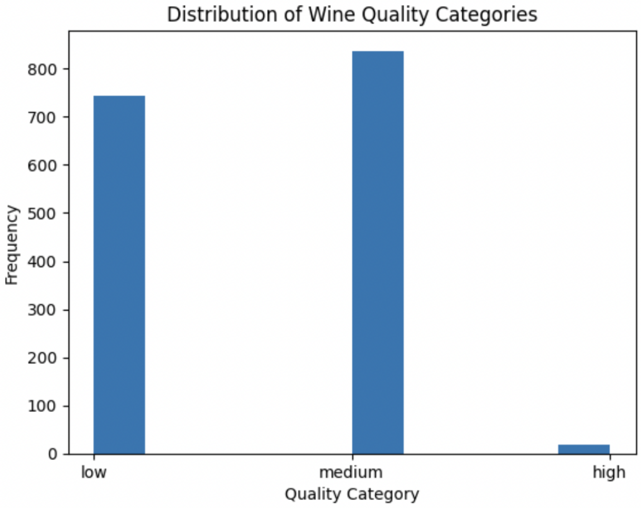


Figure 3: Quality distribution of wine dataset

These variables are used for fitting and evaluating an ordinal logistic regression model on the wine quality dataset. In general, logistic regression is used for binary classification tasks where the outcome target variable has only two categories. Since, wine's quality has more than two categories including low, medium, and high, other classification method such as multinominal logistic regression has been adopted for this experiment. It is extension of logistic regression where it models the probabilities of the input data point belonging to each category relative to a reference category.

4.2 Experimental setup

Experiment 1 shows training three different logistic regression models. The first model, `l1_model`, uses L1 regularization with a multinomial loss function and the Saga solver to fit the data. The second model, `l2_model`, uses L2 regularization with the same multinomial loss function and the Newton-CG solver to fit the data. Finally, the third model, `elastic_net_model`, uses elastic net regularization, which is a combination of L1 and L2 regularization with a 50/50 ratio between the two. This model also uses the multinomial loss function and the Saga solver to fit the data. All three models are trained using the same training data, `X_train`, and target values, `target_train`, with a maximum of 100 iterations.

Two ordinal regression models, probit and logit, are trained using OrderedModel from statsmodels. The exog (independent variables) and endog (target variable) are passed to the model, and the fit() method is used to train the model with a specific optimization method ('bfgs') The trained ordinal regression models are used to predict the target variable `target_ordinal` for the testing set `ordinal_test`. The predicted values are then converted to integer labels using `argmax` method, and the accuracy is calculated by comparing the predicted labels with the true labels `target_ordinal`.

4.3 Experiment 1

As per scikit-learn, the multinomial option is supported only by the 'lbfgs', 'sag', 'saga', and 'newton-cg' solvers Pedregosa et al. (2011). Since the wine dataset is small, the 'liblinear' solver is a good choice, while 'sag' and 'saga' solvers are faster for larger datasets. Unfortunately, the wine quality is represented by multiple categories, and 'liblinear' does not support a multinomial backend Pedregosa et al. (2011).

The "saga" solver in scikit-learn is specifically designed for large datasets, but it can also be effective for small datasets. It is an efficient solver that supports both L1 and L2 penalties, and it can handle multinomial logistic regression with L1 regularization effectively. "saga" is a variant of stochastic gradient descent (SGD) that combines the best of batch gradient descent and stochastic gradient descent, making it suitable for small datasets as it updates the model parameters incrementally based on subsets of the data. "lbfgs" is a solver that is commonly used for logistic regression with L2 penalty in scikit-learn. It stands for Limited-memory Broyden-Fletcher-Goldfarb-Shanno, which is a type of quasi-Newton optimization algorithm.

To increase accuracy and performance of the model, k-fold cross-validation with five iteration on training data have been performed. Since, **sulphur dioxide** is highly deviated, it undergoes `min-max` scaling for normalizatin. Overall, "lbfgs" is a popular solver for logistic regression because it is fast, memory-efficient, and performs well on a wide range of datasets. For small to medium-sized datasets and commonly used solvers such as `"newton-cg"`, `"lbfgs"`, `"liblinear"`, and `"sag"`, the default value of `max_iter=100` is often sufficient. These solvers typically converge relatively quickly, and increasing the number of iterations beyond what is necessary may not result in significant improvements in model performance.

4.4 Experiment 2

For ordinal regression, exogenous variable are same as pervious one but target variable is ordered wine quality categories including 3, 4, 5, 6, 7 and 8. Probit and oridnal regression does not perform well as shown in Table 1 if we use wine quality as orginal. It is due to lack of data frequency. Therfore, wine quality are further categoized into low, medium and high. Rating less than 6 are considered ”low” where as more than 7 are labled as ”high”

An ordinal logistic regression model is created using the OrderedModel class, with `targetordinal_train` as the predictor variable and `ordinal_train` as the ordinal response variable. The model assumes a probit distribution. The fit method is then used with the BFGS optimization algorithm to estimate the model parameters. Summary show that it was successfully optimized using the BFGS algorithm. The current function value, which represents the value of the objective function at the final converged point, was found to be 0.973498. The optimization process required 33 iterations, with a total of 34 function evaluations and 34 gradient evaluations. The model's parameters were estimated with a relatively small number of iterations and function evaluations.

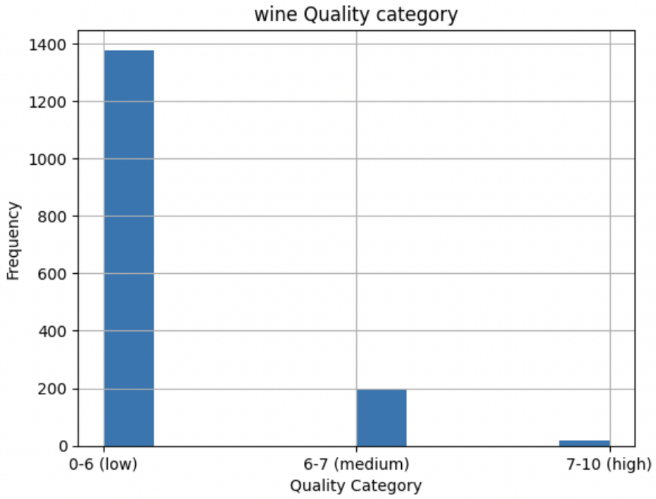
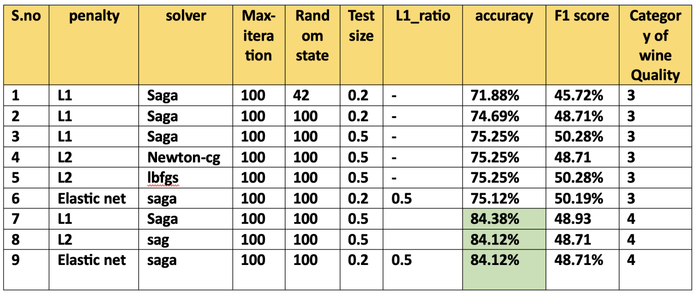


Figure 4: further category of wine quality

4.5 Results

Following Figure 5 is the result generated from **Experiment 1** which is of scikit-learn multinomial target. As seen from the given figure, best accuracy prediction is achieved when using L1 penalty with saga solver and test size of 0.5. All L1, l2 and elastic net give accuracy result with 70's range when quality of wine is categorised into 3 label. Apparently, it shows that accuracy increased when wine quality's target is divided into four categorical.But F1 score is not increased.In Conclusion, it can be said that test data accuracy prediction depends upon multiple factor including its test size, random state and dataset standardization.



S.no	penalty	solver	Max-iteration	Random state	Test size	L1_ratio	accuracy	F1 score	Category of wine Quality
1	L1	Saga	100	42	0.2	-	71.88%	45.72%	3
2	L1	Saga	100	100	0.2	-	74.69%	48.71%	3
3	L1	Saga	100	100	0.5	-	75.25%	50.28%	3
4	L2	Newton-cg	100	100	0.5	-	75.25%	48.71	3
5	L2	lbfgs	100	100	0.5	-	75.25%	50.28%	3
6	Elastic net	saga	100	100	0.2	0.5	75.12%	50.19%	3
7	L1	Saga	100	100	0.5		84.38%	48.93	4
8	L2	sag	100	100	0.5		84.12%	48.71	4
9	Elastic net	saga	100	100	0.2	0.5	84.12%	48.71%	4

Figure 5: sklearn logestic Regression's accuracy Results

Table 1 summarises results is of statsmodel's ordinal regression and that discussed in Experiment 2.

Table 1: Accuracy of methods using different distributions

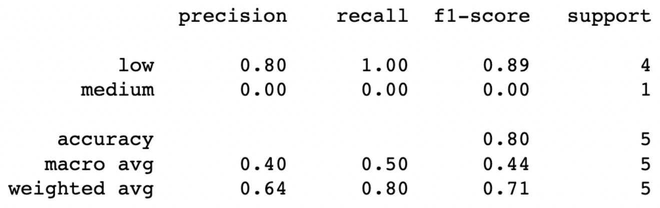
Distribution	Method	Test Size	Accuracy	category
Probit	BFGS	0.2	63	6
Logit	BFGS	0.2	62	6
Probit	BFGS	0.5	60	6
Logit	BFGS	0.5	61	6
Probit	BFGS	0.2	84.64	3
Logit	BFGS	0.2	84.64	3

Probit and logit Distribution with BFGS method with test size of 0.2 and categoized into ”low”, ”medium” and ”high” tends to perform better with accuracy fraction of 84.64 percentage. It cannot be said that model is better fit solely based on accuracy. By examining the distribution of the errors or residuals using techniques such as histograms, density plots, or other visualizations, it can be assess whether the errors in the data follow a normal distribution, logistic distribution, or some other distribution. This will help to determine which distributional assumption is more appropriate for the data and choose the appropriate model accordingly. Probit regression is preferred when the data with normally distributed errors or a symmetric error distribution. Probit models are also commonly used in situations where the ordinal outcome represents a probability or a proportion.

As per wine test data, the residuals of both the probit and logit models exhibit similar characteristics, such as being normally distributed and symmetric, it may suggest that both models are performing similarly well in terms of capturing the underlying patterns in the data. In such cases, other factors such as model interpretability, ease of implementation, or specific requirements of the analysis or application may be considered in selecting between probit and logit.

5 Discussion

Five raw data had been separated which do not undergoes training and testing of scikit and statsModels Logistic regression. Now, five raw data had been fitted to different kind of model to predict its performance.Overall, the model has a high precision and recall for the "low" class, but it performs poorly for the "medium" class. The weighted average F1-score is 0.89, indicating good overall performance, but it should be noted that the "medium" class has no predicted instances, resulting in low metrics for that class.There are many reasons for the poor performance on the "medium" class and can improve the model accordingly.



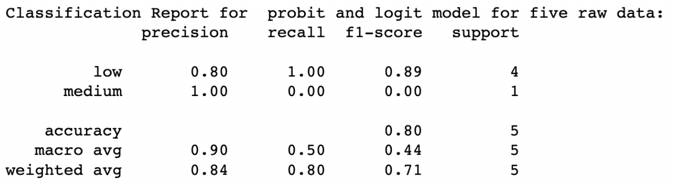
```
              precision    recall  f1-score   support

         low       0.80      1.00      0.89         4
      medium       0.00      0.00      0.00         1

    accuracy                           0.80         5
   macro avg       0.40      0.50      0.44         5
weighted avg       0.64      0.80      0.71         5
```

Figure 6: L1 model prediction on five Raw data

One reason is that low wine Quality have frequency more than 1200 frequency in training set where as others have less than 200. accuracy of prediction directly proportional to frequency of taining dataset.



```
Classification Report for probit and logit model for five raw data:
              precision    recall  f1-score   support

         low       0.80      1.00      0.89         4
      medium       1.00      0.00      0.00         1

    accuracy                           0.80         5
   macro avg       0.90      0.50      0.44         5
weighted avg       0.84      0.80      0.71         5
```

Figure 7: Probit model prediction on five Raw data

As clearly seen from classification report, precision on five raw data is 100 percent for low quality while it often confused with high and medium quality.

6 Conclusion

Both sklearn and stats model are popular Python libraries for machine learning and statistical analysis. sklearn provides a wide range of machine learning algorithms and tools for tasks such as classification, regression, clustering, and model evaluation, while stats model focuses more on statistical models for data analysis, including linear regression, logistic regression, and ANOVA. The performance of a model depends on various factors, including the specific algorithm used, hyperparameter tuning, feature engineering, and dataset characteristics Based on the evaluation of the wine quality dataset using sklearn's L1, L2 and elastic net , it can be concluded that the model's performance is better for the "low" class compared to the "medium" class. The precision, recall, and F1-score for the "low" class are relatively high, indicating that the model is able to correctly predict the "low" class instances

References

Chen, J. (2019, May 1). 5 types of regression and their properties. *Towards Data Science*. Retrieved from `https://towardsdatascience.com/5-types-of-regression-and-their-properties-c5e1fa12d55e`

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization path for generalized linear models by coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Grace-Martin, K. (2018). *The difference between logistic and probit regression.* `https://www.theanalysisfactor.com/the-difference-between-logistic-and-probit-regression/`. (Accessed on April 14, 2023)

IBM. (2021). *Ordinal Regression.* `https://www.ibm.com/docs/el/spss-statistics/25.0.0?topic=features-ordinal-regression`. ([Online; accessed 14-April-2023])

Kim, S. J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*(4), 606–617.

Minka, T. P. (2003). A comparison of numerical optimizers for logistic regression. In *Advances in neural information processing systems* (pp. 865–872).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Simon, N., Friedman, J., & Hastie, T. (2011). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *Statistical computing and graphics newsletter*, *22*(2), 26–33.