Develop Modeling Report 3 Version 2

The modeling approach employed in this report revolves around expanding regression models and adjusting hyperparameters in order to simulate logistic regression. Building upon insights from the previous week's analysis, the focus lies on enhancing the accuracy of predicting monetary policy action and its degree. This model will follow the example of previous weeks by taking the fitted predicted values from the models and converting them into the closest monetary policy action level (from -1 to 1 in steps of 25 basis points). More advanced regression models will generate accurate estimations of the degree of monetary policy as well as the direction.

1. Ordinal Ridge Regression

Ordinal Ridge Regression presents significant benefits in predicting the degree and direction of monetary policy action. The approach refines logistic regression models, enhancing the accuracy of determining monetary policy action. Firstly, it maintains the ordinal direction of increasing, decreasing, or holding, since they are not just words, but have information encoded. Ordinal Ridge Regression takes this into consideration when building the model. The model also benefits from estimating probabilities for each category rather than just class predictions. Thus, it provides more information about the predicted values for each observation.

Metric	Training Set	Validation Set
MSE	0.0373037	0.0729167
RMSE	0.193142	0.270031
R^2	0.0731313	-0.12187
Adjusted R^2	-0.834428	1.36342
Accuracy	0.706806	0.583333

This model does not perform as well enough to be considered usable in the test set. An R^2 of -0.122 in the validation set do not meet the requirements for use beyond model selection, though it will provide a healthy baseline accuracy metric on which to evaluate the following models. The accuracy of 0.583 in the validation set is also a good starting point to evaluate further models on.

2. XGBoost Model

XGBoost Models also provide benefits when predicting multi-level ordinal logistic variables like the 'Difference' variable. Firstly, boosting is a helpful method of combining predictions of multiple weak learners to create a strong predictive model. It also allows for feature importance, showing which variables are the most integral to predicting the outcome variable. This will come in handy during hyperparameter selection. Finally, it also performs regularization, which helps avoid overfitting.

These models are significantly more complex than previous models, but offer that much more in accuracy. For example, XGBoost uses gradient boosting to optimize the model's performance. Gradient boosting involves adjusting the weights of each training instance based on the errors made by the previous model. It essentially gives more importance to the data points that the previous model predicted incorrectly.

XGBoost Model - All Variables

Metric	Training Set	Validation Set
MSE	0	0.0208333
RMSE	0	0.144338
R^2	1	0.679466
Adjusted R^2	1	1.10384
Accuracy	1	0.666667

This model performs suspiciously well in the training set, but is weaker when generalized to the validation set. This shows that XGBoost Models do add some predictive power over Ordinal Ridge Regression thanks to its many features, but perhaps due to overfitting. The accuracy score also greatly increases confidence in this model. From here, hyperparameter selection will help avoid overfitting while ideally retaining the much higher accuracy.

XGBoost Model - Reverse Variable Selection

Metric	Training Set	Validation Set
MSE	0	0.0208333
RMSE	0	0.144338
R^2	1	0.679466
Adjusted R^2	1	1.10384
Accuracy	1	0.666667

Reverse variable selection was used to simplify the model as well as reduce the possibility of overfitting. This is an automated way of reducing dimensionality without reducing accuracy. This model shows no difference in accuracy, which is interesting because it has 98 fewer variables than the previous model. This just goes to show how many variables were included, and how overfit they were.

XGBoost Model – Manual Variable Selection

Metric	Training Set	Validation Set
MSE	0	0.046875
RMSE	0	0.216506
R^2	1	0.278798
Adjusted R^2	1	1.23363
Accuracy	1	0.375

One of the benefits of XGBoost Models is that they show feature importance for all the variables. This offers some methods of variable selection. Here, all variables with greater than 0.01 feature importance were chosen for a subset of independent variables for the sake of simplicity and interpretability, resulting in 18 variables—much less than the previous two models. It produces admirably high accuracy scores in the training set, but suffers heavily when generalized to the validation set. This proves some balance in bias between overfitting and simplicity. Given that the training there is only upside to the increased model complexity, the reverse variable selection seems to be the best model within the XGBoost family.