# Develop Modeling Report 2 Version 2

The modeling approach employed in this analysis involves Random Forest Models and Support Vector Models. These models increase complexity relative to the previous week's OLS models. These models allow for non-linearity where OLS models are strictly linear. This allows for more complex models with greater predictive power. They are also both ensemble learning techniques, which combine the predictions of multiple models. It is also robust to outliers, where OLS models are not.

Like the previous analysis, there is also some post-processing occurring, as linear models will produce an array of fitted values that are not always in the form of monetary policy actions necessary. As such, all fitted values are converted or "rounded" to its nearest monetary policy action-accepted value.

1. Random Forest Classifier

A Random Forest Classifier is another method of multi-level logistic regression. The Random Forest Classifier offers several benefits in the context of machine learning. It is an ensemble learning method that combines multiple decision trees to make predictions, resulting in a robust and accurate model. One key advantage is its ability to handle high-dimensional data with a large number of features, as it selects a subset of features at each split, reducing the risk of overfitting. Random Forests are also resistant to outliers and can handle missing data without requiring imputation. Random Forests are relatively more complex than multinomial logistic regression, as each tree has its own subset of features and samples, and involves the combining of multiple decision trees.

There are some hyperparameters that were chosen using example templates for writing this code. The first model used the maximum of 100 estimators and 42 random states allowed. This sets the baseline for the following models that will perform optimized hyperparameter selection.

```
+--------------+----------------+------------------+
| Metric       |   Training Set |   Validation Set |
+==============+================+==================+
| MSE          |     0.00294503 |        0.0182292 |
+--------------+----------------+------------------+
| RMSE         |     0.0542681  |        0.135015  |
+--------------+----------------+------------------+
| R^2          |     0.926826   |        0.719533  |
+--------------+----------------+------------------+
| Adjusted R^2 |     0.855177   |        1.09086   |
+--------------+----------------+------------------+
| Accuracy     |     0.95288    |        0.708333  |
+--------------+----------------+------------------+
```

This model performs very admirably, with a training accuracy of 0.953 and a validation accuracy of 0.708. This shows strength both in hyperparameter selection and generalization where previous OLS models failed. It still shows an Adjusted $R^2$ greater than 1, which indicates that there are more predictors than estimated values. This will hopefully be taken out by dimensionality reduction.

The next model attempts to select an optimal number of estimators, depth, and samples over which to predict. This is done through 5-fold cross validation to select the best Random Forest model. This model attempts to reduce dimensionality while maintaining or even increasing accuracy.

```
+--------------+----------------+------------------+
| Metric       |   Training Set |   Validation Set |
+==============+================+==================+
| MSE          |     0.00294503 |        0.0182292 |
+--------------+----------------+------------------+
| RMSE         |     0.0542681  |        0.135015  |
+--------------+----------------+------------------+
| R^2          |     0.926826   |        0.719533  |
+--------------+----------------+------------------+
| Adjusted R^2 |     0.855177   |        1.09086   |
+--------------+----------------+------------------+
| Accuracy     |     0.95288    |        0.708333  |
+--------------+----------------+------------------+
```

It seems that this model was able to reduce the dimensionality of the model without doing anything to the predictive power. This is evident from the fact that only 50 variables were selected as the maximum. That said, given that the accuracy measures are exactly the same, analysis will move on.

Finally, an optimization code to select the best value for cross validation, which would be fed into the selection process for hyperparameter selection. This altered selection process would more accurately choose the best hyperparameters.

```
+--------------+----------------+------------------+
| Metric       |   Training Set |   Validation Set |
+==============+================+==================+
| MSE          |     0.00359948 |        0.0234375 |
+--------------+----------------+------------------+
| RMSE         |     0.0599956  |        0.153093  |
+--------------+----------------+------------------+
| R^2          |     0.910565   |        0.639399  |
+--------------+----------------+------------------+
| Adjusted R^2 |     0.822994   |        1.11681   |
+--------------+----------------+------------------+
| Accuracy     |     0.942408   |        0.75      |
+--------------+----------------+------------------+
```

This model helps balance the bias toward generalization against the validation group without sacrificing the accuracy in the training set too much. It shows a strong training set validation of 0.942 and validation accuracy of 0.75. So far, this model performs the best when considering both training and validation accuracy together.

2. Support Vector Machine

Support Vector Machines (SVM) are a powerful class of supervised machine learning algorithms used for classification and regression tasks. The SVM model is trained with a linear kernel, which allows it to learn linear decision boundaries between different classes. The accuracy of the trained SVM model on the training data is calculated and reported. SVMs have the advantage of being effective in high-dimensional spaces and are known for their ability to handle complex datasets. They aim to find an optimal hyperplane that maximally separates different classes, thus making them a valuable tool for various classification problems.

Three different models were chosen for Support Vector Machines, optimizing over the model parameter and cross validation fold numbers. 3-, 5-, and 7-fold cross validation was used to select the best model family and hyperparameters, and they all returned the same model with the following accuracy measures:

```
+--------------+----------------+------------------+
| Metric       |   Training Set |   Validation Set |
+==============+================+==================+
| MSE          |      0.0405759 |       0.0651042  |
+--------------+----------------+------------------+
| RMSE         |      0.201435  |       0.255155   |
+--------------+----------------+------------------+
| R^2          |     -0.008173  |      -0.00166945 |
+--------------+----------------+------------------+
| Adjusted R^2 |     -0.995342  |       1.32448    |
+--------------+----------------+------------------+
| Accuracy     |      0.696335  |       0.583333   |
+--------------+----------------+------------------+
```

These models do not perform as well as the previous ones in both the training and validation accuracy metrics. It also still shows an adjusted $R^2$ higher than 1, suggesting that it has much higher model complexity than required. Finally, the fact that all three optimized hyperparameter selection methods produced the same model suggests that it is not worth considering this class of models for the best model.