Project Report- Forecasting U.S Mortgage Rates: 2024 and Beyond

Group 59: Pallavi Singh, Themiya Dias Chandraratna, Abhi Patel, Roshni Patnayakuni, Nagendra Achanta

Git Hub Repository: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-59

Scheller College of Business, Georgia Institute of Technology

MGT 6203: Data Analytics for Business

Dr. Jonathan Fan

December 1, 2023

Table of Contents

Project Background

In the past 3 years, interest rates have more than doubled from roughly 3% for a 30-year fixed mortgage to well over 7%. October 2023 saw mortgage rates hit 7.79% which is the highest average 30-year mortgage rate since November 2000 (Rothstein, 2023). Mortgage rates are a key driver of the housing market and Central Bank rates are a prime factor affecting movements in mortgage rates (Sadorsky, 2015, p.822). Monetary policy changes affect market rates of interest because of federal fund rate pass-through (Hegwood & Tuttle, 2017, p.57). Although monetary policy changes affect market rates of interest, other factors can play a significant role (Arena et al., 2020, p.11). For example, Kliesen and Schmid (2004) found a statistically significant response of inflation expectations to surprises in the CPI, the core CPI, retail sales, and the NAPM index (p.10). Overall, the difference between inflation expectations and actual inflation tends to narrow (Kuncoro, 2020, p.76). This study aims to forecast U.S 30-year fixed mortgage rates based on macroeconomic data components.

Purpose

Interest rates have a significant effect on the economy, influencing consumer spending and manufacturing rates. Mortgage rates comprise a significant cost for housing buyers; a 4% increase in a 30-year fixed mortgage interest rate, from 3% to 7%, adds $280,740.62 to the cost of a $400,000 mortgage. Lower interest rates make homes more affordable, stimulating the housing market and encouraging new buyers. Existing homeowners benefit from refinancing at lower rates, saving thousands over the loan period. Lower mortgage rates increase disposable income, driving higher consumer spending on various goods and services, benefiting the broader economy. Higher rates deter potential buyers due to increased borrowing costs and may cool housing market activity. Rate fluctuations can cause financial stress for homeowners, especially those with adjustable-rate mortgages. Accurate mortgage rate forecasts are valuable for investors, firms, government entities, and homebuyers. The models and techniques from this study could potentially provide accurate mortgage rate forecasts and trends for 2024 and beyond.

Data Preparation

An in-depth analysis of Federal Reserve Economic Data (FRED) was performed, focusing on key U.S. economic categories: Money, Banking, & Finance, Population, Employment, & Labor Markets, National Accounts, and Production & Business Activity. Discontinued data sources were avoided, and calculations reflecting current best practices were prioritized. Selected variables included Consumer Price Index (CPI) for all items, Total Federal Public Debt, Households Net Worth, Housing Inventory Active Listing Count,

Industrial Production Consumer Goods, Industrial Production Total Index, 3-Month Interest Rates, 10-Year Treasury Yield, Job Openings Total in non-farming, M2 money supply, Producer Price Index (PPI) for all commodities, Real Gross Domestic Product (GDP), Unemployment Rate, and Velocity of M2 money supply.

Datasets were gathered from FRED and stored in a MySQL relational database on Amazon Web Services (AWS) Relational Database Service (RDS). Data was manipulated to convert record dates to Year-Month format (%Y-%m) for consistent matching. Initially, joining all selected factors by record date yielded only 28 records for 14 predictors, falling short of the recommended 10 records per predictor variable for regression analysis. Further joins were explored, leading to a dataset comprising 8 predictors (Industrial Production Consumer Goods, CPI, Households Net Worth, 3-Month Interest Rates, Job Openings Total in non-farming, M2 money supply, and Velocity of M2 money) with the response variable of 30-Year Fixed Rate Mortgage Average. To handle multiple records per month for mortgage rates and money supply, average monthly values were used after additional data manipulation. After conducting exploratory data analysis to assess the reliability of the selected predictors, a Java code was developed to calculate period lag on mortgage rates, enabling the computation of forecast horizons spanning 3, 6, and 12 months. This step enhanced the predictive capability of the dataset.

Handling Null values

In the data preparation process and leveraging inner joins, the dataset was cleared of null values and the time-period and dataset were created such that any factors or fields that proved to contain nulls were avoided. Imputation can be leveraged to fill null values and regression can be used on predictor variables to provide accurate values.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial initial phase following data collection and pre-processing. During this stage, the data is visualized and manipulated without any underlying assumptions. This process aids in assessing the data's quality and serves as a foundation for model building. To examine correlations between attributes, a correlation matrix was plotted, revealing the correlation coefficients between the variables.

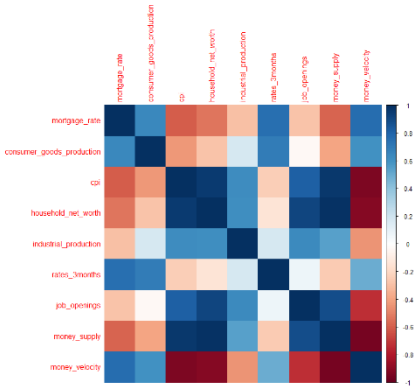

Fig.1. Correlation Matrix Heatmap

From the correlation matrix above [Fig.1], high correlation between .8 and 1 is observed between velocity with money_supply, cpi, and household_net_worth.

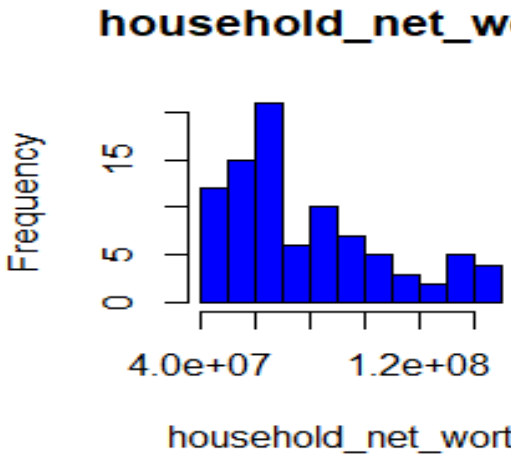

Fig.2. Histogram of Household Net Worth

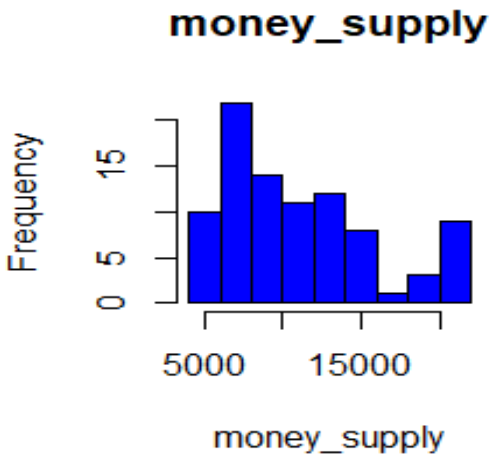

Fig.3. Histogram of Money Supply

Histograms are among the most useful EDA techniques, and allow you to gain insights into your data, including distribution, central tendency, spread, modality and outliers (Komorowski et al., 2016). We created histograms for all variables, revealing skewedness in most cases (all plots not shown here). These variables tend to deviate from a normal distribution. Logarithmic transformation can be applied to address this skewedness.

The below figure shows pair plots for the mortgage rate and all the other numeric variables. From the graphs, we can see that variables like CPI and household net worth are directly correlated. Variables like mortgage rate, industrial production, and rates_3 months are scattered randomly and hence not

correlated. We can see similar observations for CPI vs household net worth and mortgage rate vs industrial production.

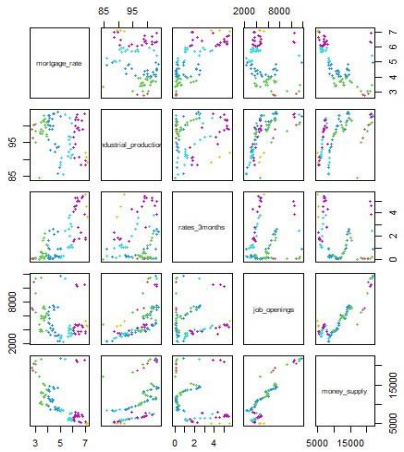

Fig.4. Pair Plot of Mortgage Rate and Predictors

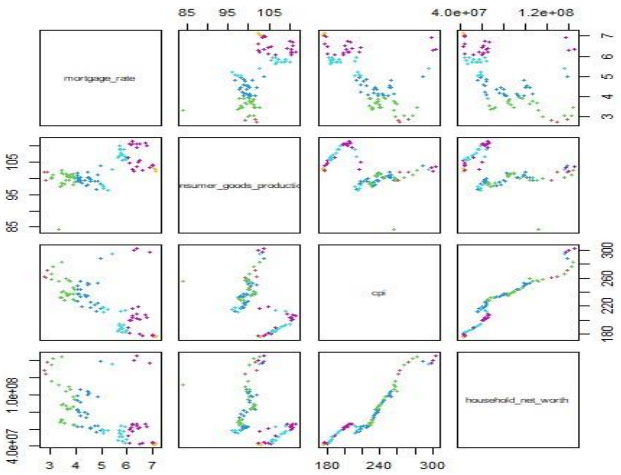

Fig.5. Pair Plot of Mortgage Rate and Predictors

Feature Selection

Feature selection is the process of reducing variables to develop a simple and efficient model. It is an important step to improve the cost of computation. Many models, especially those based on regression slopes and intercepts, will estimate parameters for every term in the model. Because of this, the presence of non-informative variables can add uncertainty to the predictions and reduce the overall effectiveness of the model (Brownlee, 2019).

Pearson coefficient - Pearson's Method is one of the most common techniques used for a linear correlation with numerical variable input and numerical variable output. Target variable is mortgage rate and from heatmap, we find out strong and weak correlation with independent variable and set threshold. The threshold of 0.7 is set to find highly correlated variables. Household_net_worth, job openings, money supply, money velocity are highly correlated, and few can be dropped to reduce dimensionality.

Analytical Approach

Initially, a regression model was trained on mortgage rates and 8 selected predictors to assess reliability. The model showed high significance for several predictors, resulting in an adjusted R^2 value of 0.8748, signifying a good fit. Following Sadorsky's (2015) approach, lag periods of 3, 6, and 12 months were used for forecasting mortgage rates. The 1-month period was omitted due to its limited relevance in the context of financial quarters (p. 823).

```
Call:
lm(formula = mortgage_rate ~ ., data = corr_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84352 -0.31655  0.00891  0.29326  1.17226

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -7.079e+00  3.362e+00  -2.105   0.0383 *
consumer_goods_production -4.494e-03  3.035e-02  -0.148   0.8826
cpi                       7.465e-03  1.157e-02   0.645   0.5206
household_net_worth      -7.407e-08  1.778e-08  -4.166 7.71e-05 ***
industrial_production    -4.978e-02  2.749e-02  -1.811   0.0738 .
rates_3months             2.733e-01  6.095e-02   4.484 2.39e-05 ***
job_openings              1.118e-04  9.968e-05   1.122   0.2652
money_supply              6.790e-04  2.185e-04   3.107   0.0026 **
money_velocity            7.813e+00  1.865e+00   4.190 7.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4391 on 81 degrees of freedom
Multiple R-squared:  0.886,     Adjusted R-squared:  0.8748
F-statistic: 78.72 on 8 and 81 DF,  p-value: < 2.2e-16
```

Fig. 6 Initial Regression Model

```
> summary(model_3_months)

Call:
lm(formula = X3monthForecast ~ consumer_goods_production + cpi +
    household_net_worth + industrial_production + rates_3months +
    job_openings + money_supply + money_velocity, data = file_path)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92570 -0.26591  0.01864  0.30567  1.06745

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.371e+01  3.653e+00  -3.752 0.000329 ***
consumer_goods_production  3.361e-03  3.297e-02   0.102 0.919055
cpi                        1.121e-02  1.257e-02   0.892 0.375255
household_net_worth       -6.553e-08  1.932e-08  -3.391 0.001077 **
industrial_production     -5.521e-02  2.987e-02  -1.849 0.068154 .
rates_3months              1.131e-01  6.623e-02   1.707 0.091677 .
job_openings               1.738e-04  1.083e-04   1.605 0.112451
money_supply               7.327e-04  2.374e-04   3.086 0.002774 **
money_velocity             1.035e+01  2.026e+00   5.108 2.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4771 on 81 degrees of freedom
Multiple R-squared:  0.8647,    Adjusted R-squared:  0.8513
F-statistic: 64.69 on 8 and 81 DF,  p-value: < 2.2e-16
```

Fig.7. Linear Regression Model for 3month Forecast

The 3-month regression model (Fig. 6) demonstrates a significant relationship between household_net_worth, rates_3months, and money_velocity with mortgage rates at a 99.9% confidence level. In the 6-month forecast model (Fig. 8), money_velocity remains significant at 99.9% confidence, and job_openings and industrial_production also exhibit significance. Similar patterns emerge in the 12-month forecast model (Fig. 9). These significant parameters can inform further analysis for creating a new model. However, with the increase in the forecasting period, the adjusted R^2 decreases, indicating reduced reliability of the independent variables in predicting mortgage rates. Testing through cross-validation can assess the accuracy of the models and determine the best one. Based on the fit using adjusted R^2, the regression model for 3-month forecasts proves to be the most accurate.

```
Call:
lm(formula = X6monthForecast ~ consumer_goods_production + cpi +
    household_net_worth + industrial_production + rates_3months +
    job_openings + money_supply + money_velocity, data = file_path)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24870 -0.34352  0.04104  0.32007  1.24992

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.429e+01  3.982e+00  -3.588 0.000569 ***
consumer_goods_production  3.552e-02  3.594e-02   0.988 0.325858
cpi                        2.200e-02  1.370e-02   1.606 0.112273
household_net_worth       -4.938e-08  2.106e-08  -2.345 0.021466 *
industrial_production     -9.416e-02  3.255e-02  -2.893 0.004902 **
rates_3months              3.215e-02  7.218e-02   0.445 0.657195
job_openings               3.279e-04  1.181e-04   2.778 0.006794 **
money_supply               5.221e-04  2.587e-04   2.018 0.046923 *
money_velocity             9.746e+00  2.208e+00   4.414 3.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.52 on 81 degrees of freedom
Multiple R-squared:  0.8418,    Adjusted R-squared:  0.8261
F-statistic: 53.86 on 8 and 81 DF,  p-value: < 2.2e-16

> |
```

Fig.8. Linear Regression Model for 6month Forecast

```
> summary(model_12_months)

Call:
lm(formula = X12monthForecast ~ consumer_goods_production + cpi +
    household_net_worth + industrial_production + rates_3months +
    job_openings + money_supply + money_velocity, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57349 -0.32744  0.01377  0.37104  1.27092

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -1.131e+01  4.529e+00  -2.498  0.01457 *
consumer_goods_production  6.207e-02  4.032e-02   1.539  0.12773
cpi                       -3.454e-03  1.611e-02  -0.214  0.83073
household_net_worth       -6.061e-08  2.406e-08  -2.519  0.01380 *
industrial_production     -8.487e-02  3.777e-02  -2.247  0.02744 *
rates_3months             -1.433e-04  8.119e-02  -0.002  0.99860
job_openings               3.529e-04  1.322e-04   2.669  0.00923 **
money_supply               7.087e-04  2.949e-04   2.403  0.01861 *
money_velocity             8.489e+00  2.489e+00   3.411  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5803 on 79 degrees of freedom
Multiple R-squared:  0.796,     Adjusted R-squared:  0.7754
F-statistic: 38.54 on 8 and 79 DF,  p-value: < 2.2e-16

> |
```

Fig.9. Linear Regression Model for 12month Forecast

MODEL SELECTION &INTERPRETATION

ARIMA

ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors. An ARIMA model is characterized by 3 terms: p, d, q. To determine optimum p,q,d for best fit model, pmdarima package provides auto_arima() which uses a stepwise approach to search multiple combinations of p,d,q and chooses the best model that has the least AIC. Our model was built using (0,2,1) for (p,q,d) as suggested by auto ARIMA. We predicted our model output and later used it to calculate the model errors (Sadorsky, P. (2015). Forecasting Canadian mortgage rates).

```
model_3m = pm.auto_arima(df["3monthForecast"], start_p=1, start_q=1,
                         test='adf',       # use adftest to find optimal 'd'
                         max_p=3, max_q=3, # maximum p and q
                         m=1,              # frequency of series
                         d=None,           # let model determine 'd'
                         seasonal=False,   # No Seasonality
                         start_P=0,
                         D=0,
                         trace=True,
                         error_action='ignore',
                         suppress_warnings=True,
                         stepwise=True)
```

Fig.10 Auto Arima timeseries forecast model

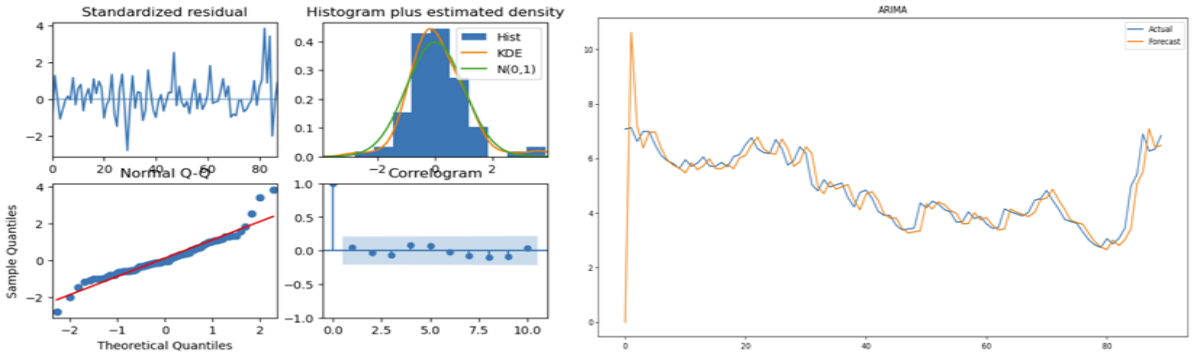

Fig.11. Diagnostic plots for standardized residuals using ARIMA

Fig.12. Actual vs Predicted ARIMA forecast model

Regression using Artificial Neural Network

Deep Learning

Albanesi and Vamossy (2019) emphasize the need for deep learning to capture the complexity of consumer default behavior, asserting that all deep learning models significantly outperform logistic regression (p.3). Given our use of similar high-dimensional data with intricate interaction patterns, standard regression might yield poor results when compared in cross-validation (Albanesi and Vamossy, 2019, p.1).

We used Neural network model in Tensor flow with Adams optimizer and mse as loss function. We used 1 hidden layer with 64 nodes and an input layer with 128 nodes. We used scaled data to build these models, plot below shows that the model fits well on both training and validation data. We examined this idea across all models built for different forecast data(3M,6M,12M) used as predictor variable. Plots below show the actual vs predicted values of our model. Errors calculated are later used to analyze the model performance (Albanesi, Stefania, and Domonkos F. Vamossy. 2019. "Predicting Consumer Default: A Deep Learning Approach").

```
model = Sequential()
model.add(Dense(128,input_dim=8,activation='relu'))
model.add(Dense(64,activation='relu'))
model.add(Dense(1,activation='linear'))
model.compile(loss='mean_squared_error', optimizer='adam',metrics=['mae'])
```

Fig.13. ANN Regression model

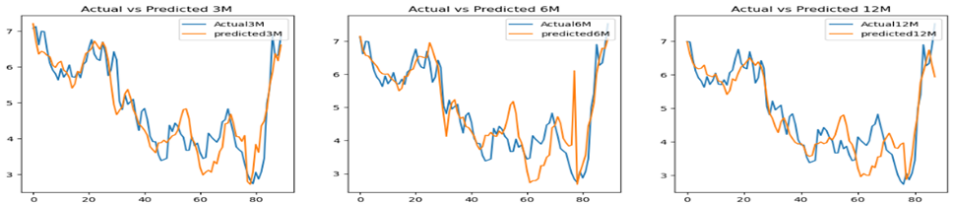

Fig.14 ANN Regression model Mortgage rate forecast (3M, 6M, 12Months)

Regression Using Decision Trees

We used DecisonTreeRegressor regression model in the scikit learn library. We built a basic tree model without using any hyper parameters, which means the model uses all the nodes and depth of the tree is unregulated. We trained model using the training dataset and predicted the output using the entire dataset. Later we used the predicted data to calculate model errors.

```
#DT regression 3m
from sklearn.tree import DecisionTreeRegressor
treee=DecisionTreeRegressor()
treee.fit(X_train_scaled,y_train)
```

Fig.15 Decision Tree Regression model

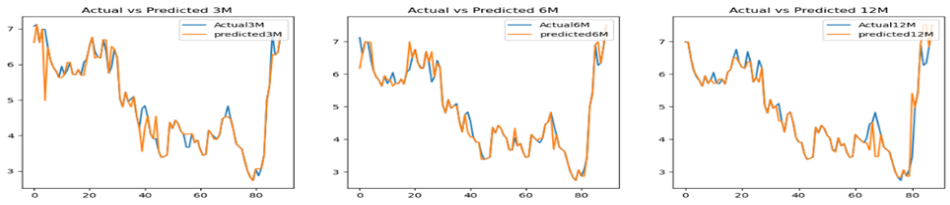

Fig.16. Decision Tree Regressor model forecast (3M, 6M, 12M)

Model Hypothesis / Interpretation:

We calculated and analyzed 6 different errors (table below) by comparing the model predicted values to actual values. All the models were performed comparatively well on both training and validation data. Decision Trees have a slight edge over the other two models, where model performed much better with lower percentage error (better accuracy, highlighted yellow) and lower overall model errors. ANN regression model performed moderately well over ARIMA forecast. Decision trees are more prone to overfitting unless carefully pruned (hyper tuning). Like in our case where we used all the features as tree nodes and no regulation of tree depth could have allowed our model overfitting. However, ANN has the advantage of learning the data again and again (epochs) has ability to conform to a better generalized model. Another drawback with our smaller sample size is that the models overperforming and there is no guarantee that they will work effectively on a new data set.

Model Error	ARIMA (Forecast)	ANN (Regression)			Decision Tree (Regression)		
		3M rate Forecast	6M rate Forecast	12M rate Forecast	3M rate Forecast	6M rate Forecast	12M rate Forecast
mape	0.078(7.8%)	0.084	0.112	0.099	0.014	0.012	0.022
me	-0.088	-0.0203	0.064	-0.047	-0.03	-0.002	-0.025
mae	0.418	0.369	0.461	0.437	0.075	0.066	0.101
mpe	-0.012	0.006	0.028	0.001	-0.004	-0.001	-0.003
mse	0.957	0.242	0.475	0.314	0.049	0.035	0.101

Table1. Forecasting models Error table

Improvements:

Analyze Training & Testing Set Sizes

Linear regression model is run by splitting the data into different test and training sizes to analyze model performance. 10%, 20%, 50% and 80% training size is tested where 10% corresponds to 60 training points and 80% is 500 training points. It is clear from graphs below that R2 improves considerably as training size increases while RSME drops significantly till 300 training set size which is 50%. After this point, RSME reduces proportionately and reaches lowest at 500.

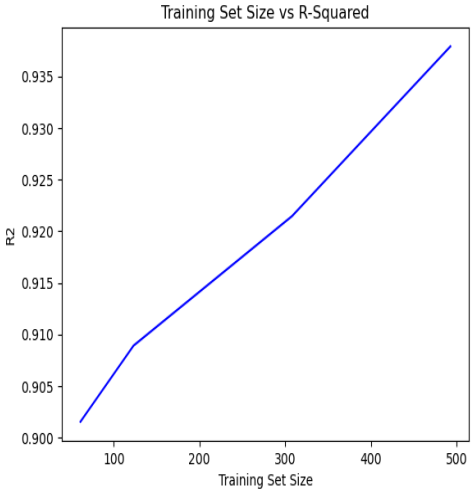

Fig.17. Training size vs R-squared

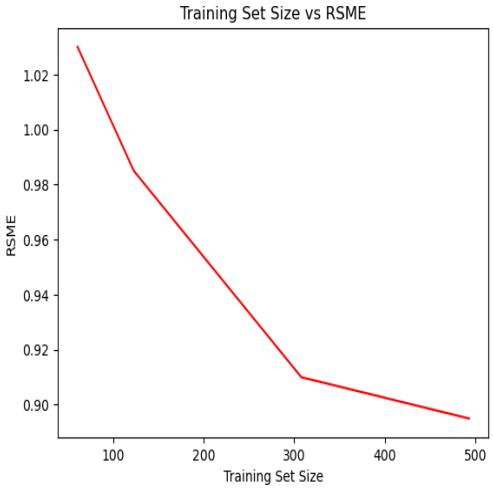

Fig.18. Training size vs RSME

Imputation of dataset

After analysis of the initial dataset, a concern arose that the array of features selected narrowed the dataset too aggressively. The original FRED data was then re-analyzed using PANDAS. A dataset was compiled with MORTGAGE30US, CPIAUCSL, INDPRO, IPCONGD, IR3TIB01USM156N, PPIACO, and UNRATE to maximize the number of datapoints while maintaining the most possible relevant features from the FRED data sources. Training sets of roughly 60, 120, 300, and 500 were created. As a frame of reference, 60 datapoints is equivalent to the initial dataset that was used for analysis. The trends of adjusted R^2 and RSME showed that more datapoints would reduce the error and increase the overall accuracy of trained regression models.

Following the initial dataset analysis, a concern arose regarding the overly aggressive feature selection approach. Consequently, the original FRED dataset underwent re-analysis utilizing the PANDAS framework. This re-analysis led to the compilation of a dataset comprising MORTGAGE30US, CPIAUCSL, INDPRO, IPCONGD, IR3TIB01USM156N, PPIACO, and UNRATE, aimed at maximizing data points while retaining the most relevant features from the FRED data sources. Subsequently, training sets of approximately 60, 120, 300, and 500 data points were generated. To provide context, it is worth noting that 60 data points equated to the size of the initial dataset used in the initial analysis. Analysis of adjusted R2 and RSME trends demonstrated that incorporating more data points reduced errors and enhanced the overall accuracy of the trained regression models.

To add additional relevant features an imputation was attempted. Imputation reduced the R2 and increased the RSME.

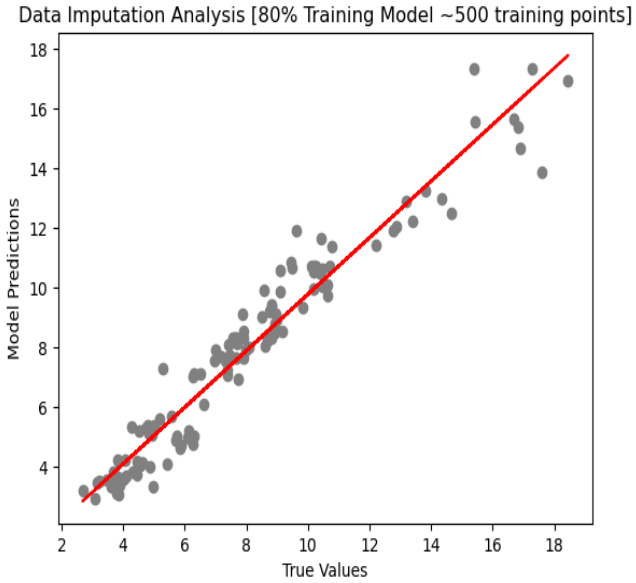

Fig.19. Data Imputation Analysis

Imputed Data:
RSME = 0.879
MSE = 0.7726191714385471
MAE = 0.6765760620264143
R2 = 0.9401435111242158

CCA Data:
RSME = 0.661
MSE = 0.43704863097311086
MAE = 0.5154023539617528
R2 = 0.9540619922552162

Fig.20. Imputed Data vs Original (CCA) Data

GitHub Repository

https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-59

Works Cited Section

Albanesi, Stefania, and Domonkos F. Vamossy. 2019. "Predicting Consumer Default: A Deep Learning Approach." National Bureau of Economic Research, working paper no. w26165, August. Available at https://doi.org/10.3386/ w26165

Arena, M., Bella, G., Cuevas, A., Gracia, B., Nguyen, V., & Pienkowski, A. (2020). It is Only Natural: Europe's Low Interest Rates (Trajectory and Drivers). *IMF Working Papers*, *2020*(116), 1–59. https://doi.org/10.5089/9781513549170.001

Brownlee, J. (2019, November 26). *How to Choose a Feature Selection Method for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Hegwood, N., & Tuttle, M. H. (2017). CONVENTIONAL MORTGAGE INTEREST RATE AND THE EFFECTIVE FEDERAL FUNDS RATE PASS-THROUGH. *Journal of Business Strategies*, *34*(1), 57–74. https://doi.org/10.54155/jbs.34.1.57-74

Kliesen, K., & Schmid, F. (2004). Monetary Policy Actions, Macroeconomic Data Releases, and Inflation Expectations. *Federal Reserve Bank of St. Louis Review*, *86*(3), 9–21. https://doi.org/10.3886/ICPSR01301

Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. (2016). Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_15

Kuncoro, H. (2020). Central Bank Communication and Policy Interest Rate. *International Journal of Financial Research*, *12*(1), 76–91. https://doi.org/10.5430/ijfr.v12n1p76

Rothstein, R. (2023, November 3). *Mortgage Rate Forecast For 2023* (C. Jennings, Ed.). Forbes Advisor; Forbes. https://www.forbes.com/advisor/mortgages/mortgage-interest-rates-forecast/

Sadorsky, P. (2015). Forecasting Canadian mortgage rates. *Applied Economics Letters*, *23*(11), 822–825. https://doi.org/10.1080/13504851.2015.1111981