Predicting Future Inflation (Final Project Report)

By: Alex Dimulescu, Benjamin Cruz, Matthew Fernandez, Reid Smith, Andrew Sullivan

Executive Summary:

The goal of this project was to see if its possible to predict future inflation in the United States (next year) utilizing factors such as interest rates and GDP in the current year, and if the prediction is found to be possible, prescribing appropriate actions consumers can take based on various potential inflation outcomes (low <2%, moderate 2-4%, and high inflation >4%). After finishing the research process, it was found that inflation can be predicted to a moderate level of accuracy (as measured by a simple accuracy measure and f1-score relative to the base accuracy). Using a logistic regression model, the simple accuracy score was 70%, and 68.1% for the f1-score (in the best case), while the base accuracy was only 42.19% (worse than a coin toss). These accuracy scores are quite good considering that the boost in accuracy from the base accuracy (always predicting low inflation <2% since it's most common) gives insight into whether the next year is going to be moderate 2-4% or high >4% inflation, which has the most potential impact on consumers. As a result it was concluded that utilizing factors such as interest rates along with GDP can predict inflation categories to an accuracy level that is sufficient enough to provide recommendations to consumers in what years to purchase goods most affected by inflation, bringing savings to the average American household.

Main Report:

Data Preprocessing:

Data Cleaning:

1. For the interest rate dataset pulled from Kaggle the Effective Federal Funds Rate (interest rate) attribute had missing values for many of the years, requiring a transformation of the original data where for years missing the interest rate the mean of the attributes values were taken and that years interest rate was set to be the computed mean (thus providing only one value for per year in the newly created data structure, the average interest rate for each year).
2. In the case of the inflation dataset pulled from Kaggle, actual inflation values (the YoY percentage change in inflation) were not provided, and instead raw CPI values were given for each month of each year. This required a transformation of the data where the mean CPI was computed for each year, and then the YoY percentage change in inflation was calculated by taking the i+1 years average CPI (i being the first year CPI data is available for) and dividing it by the ith years average CPI, incrementing i all the way to n (n being the last year CPI data was available for).

3. The GDP dataset pulled from the Federal Reserve Bank of St. Louis was already clean

Data Integration:

1. Both the interest rate and inflation datasets had non-matching start and end years, the interest rate dataset contained data from 1954-2017, while the CPI data set contained data from 1913-2022, so after both individual datasets were cleaned and the necessary attribute values were obtained for each year data was available for, a merging of the two datasets was necessary. This entails extracting a subset of the CPI data set (since the interest rate data set was the limiting factor) from 1954-2017 in order to match the dates for the interest rate data structure and then merging the two data datasets on common dates, having two features now for each date from 1954 to 2017, average interest rate and YoY percentage change in inflation.
2. The GDP dataset was obtained after the initial combination of the interest rate and CPI datasets, with its GDP ranging from 1947-2025, so the same steps as explained in part a were taken to merge GDP together with the already merged CPI and interest rate dataset.

Data Reduction:

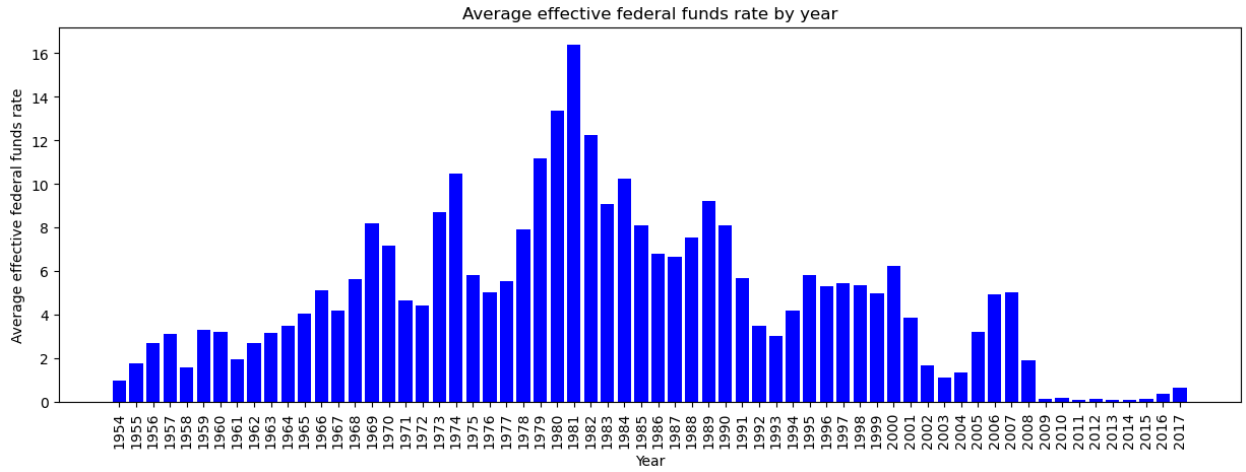
1. From the interest rate dataset factors such as Federal Funds Target Rate, Upper Target, Lower Target, Real GDP (Percent Change), Unemployment Rate, and Inflation Rate were removed. This was because inflation and GDP data were obtained from two separate datasets, and the interest rate targets were deemed redundant features since the effective rate was already being used, and unemployment was not considered useful for this analysis.

Data Transformations:

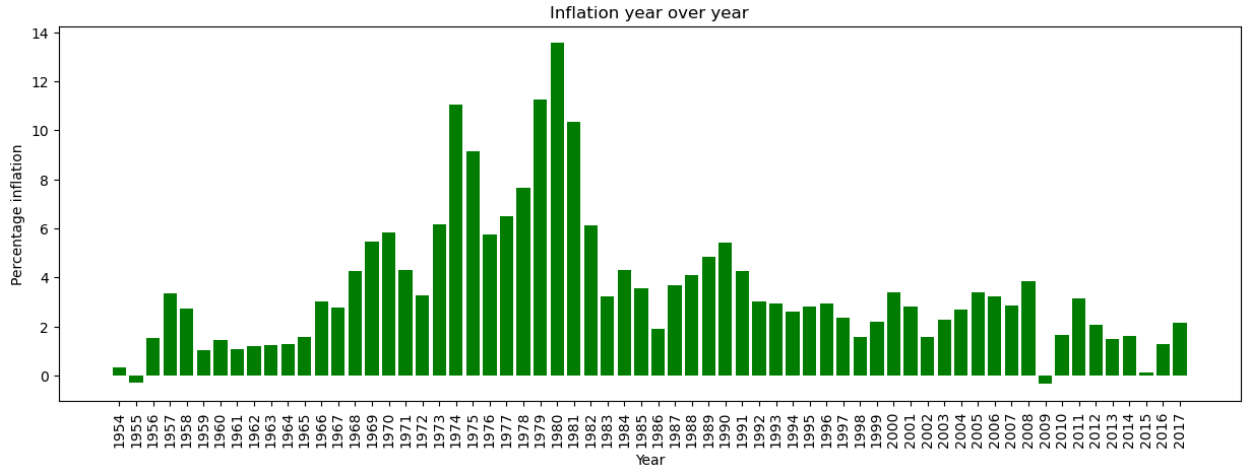
1. As mentioned, during the data cleaning step the data was technically “scaled” to the extent that the original dimensions (monthly data for interest rates and CPI) were reduced to yearly data (one data point per year for both interest rates and CPI). The GDP data fortunately was already in this same format, so no additional transformations were needed.

Exploratory Analysis of the Data:

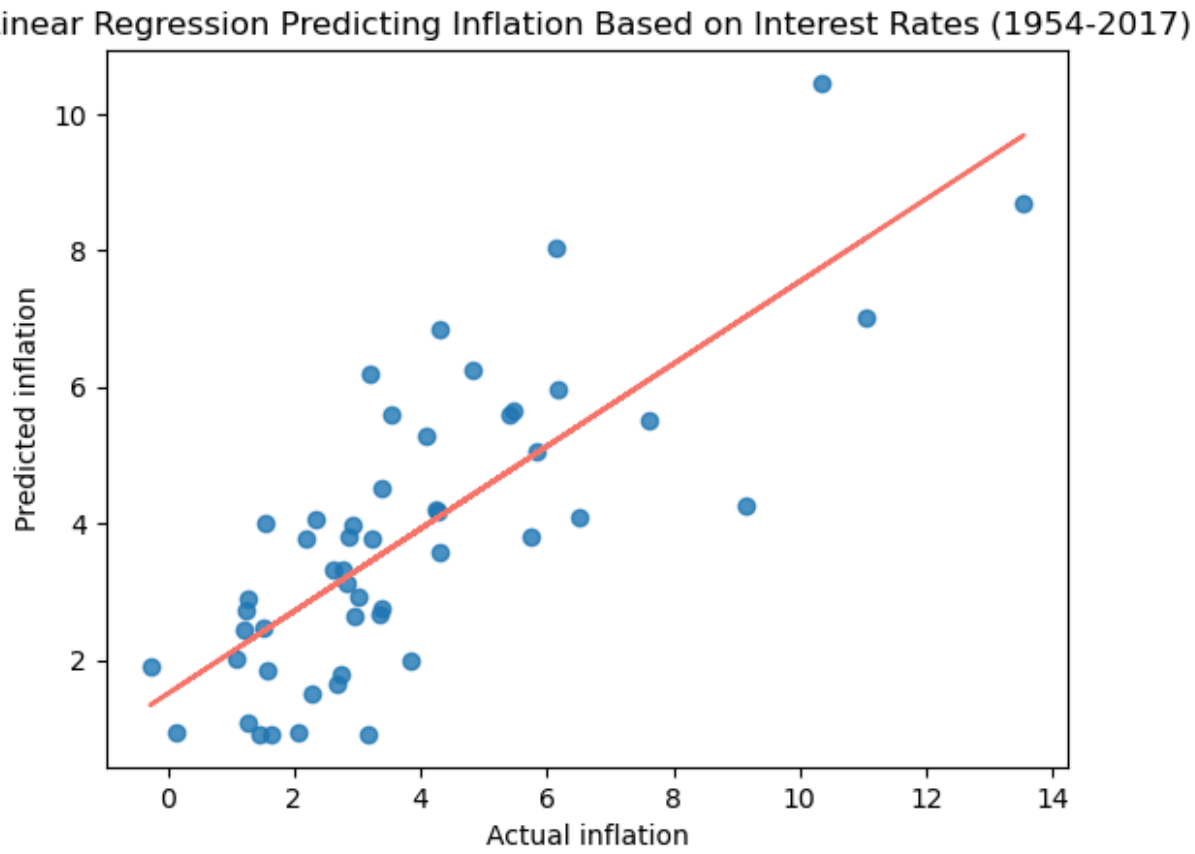
We looked into the interest rate and CPI data before beginning any actual model training to ensure that there was actually something to work with (if the data seemed altogether uncorrelated it would not make much sense to continue), and we noticed that yes in fact there seems to be quite a significant relationship between interest rates and inflation, especially in the U.S., as the FED can really only pull a few levers, the most powerful of which is interest rates. Average interest rate per year 1954-2017 histogram:



Now if you take a look at the CPI data over that same time period you get this:

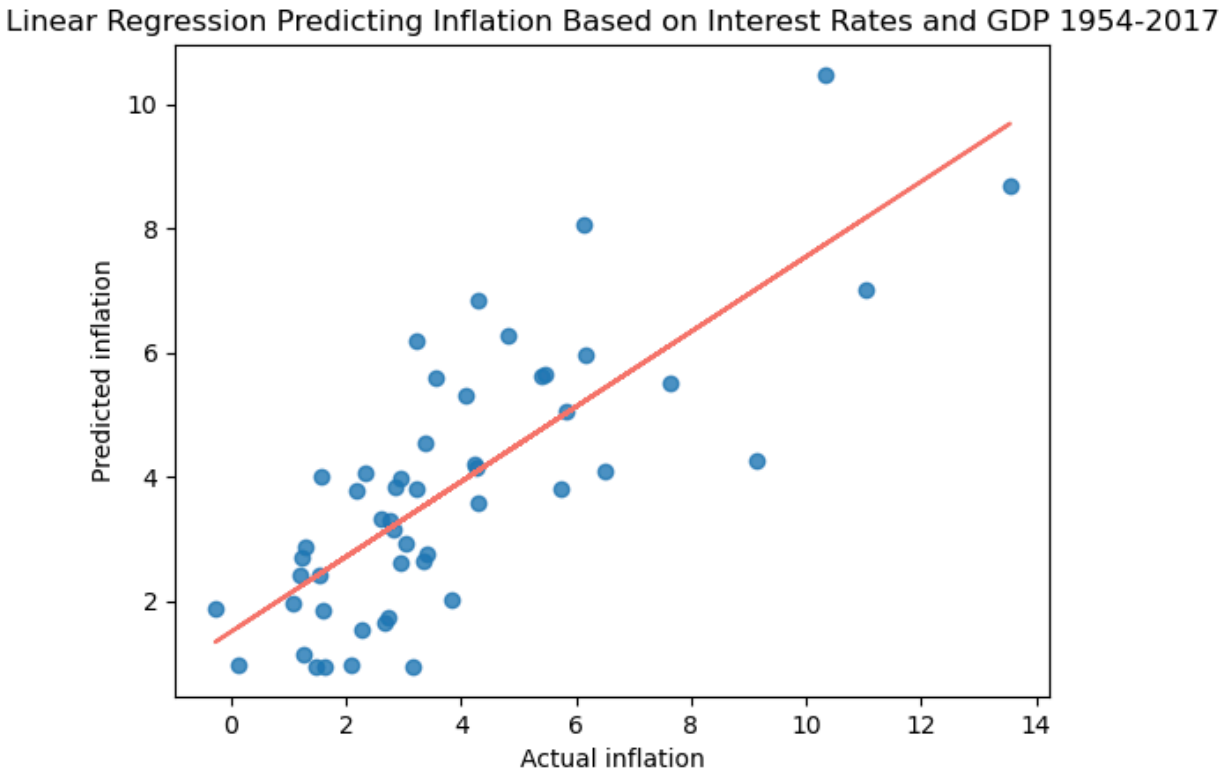


Although not identical, there is an easily identifiable relationship just from visual analysis that inflation and interest rates move in rough tandem with each other. This makes perfect sense from a macroeconomic perspective, as if inflation begins to rise (for whatever reasons), the FED can raise rates to create economic pressure in the form of higher borrowing rates, cooling the economy, and as inflation comes down, lower rates in order to restimulate economic activity. Although there are many more potential factors that go into inflation and the determination of whether to raise or lower rates, this initial relationship we saw gave us the confidence to pursue interest rates as a potential factor to predict inflation with (through the use of various modeling techniques). Now, although this was used for an early stage analysis (linear regression), we wanted to see if accuracy could be boosted (which is why we also settled on GDP, but this was after the testing of interest rates alone as a predictive factor), here is the initial linear regression and validation metrics:



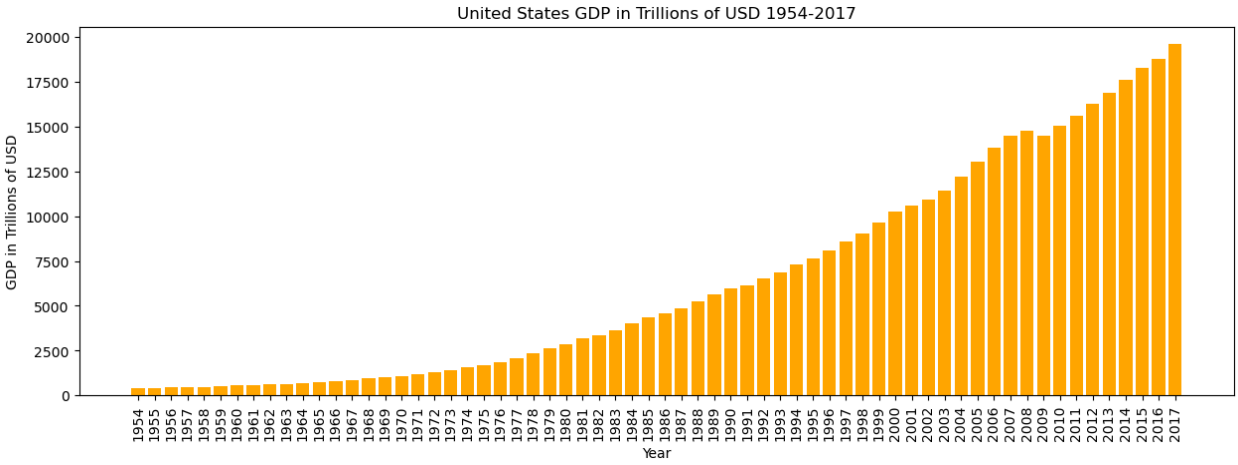
Method	Training MSE	Training R2	Test MSE	Test R2
Linear Regression	2.927423	0.603937	3.029354	0.594702

We then included GDP as a predictive factor as well and this was the resultant linear regression and validation metrics:



Method	Training MSE	Training R2	Test MSE	Test R2
Linear Regression	2.926975	0.603997	2.991599	0.599753

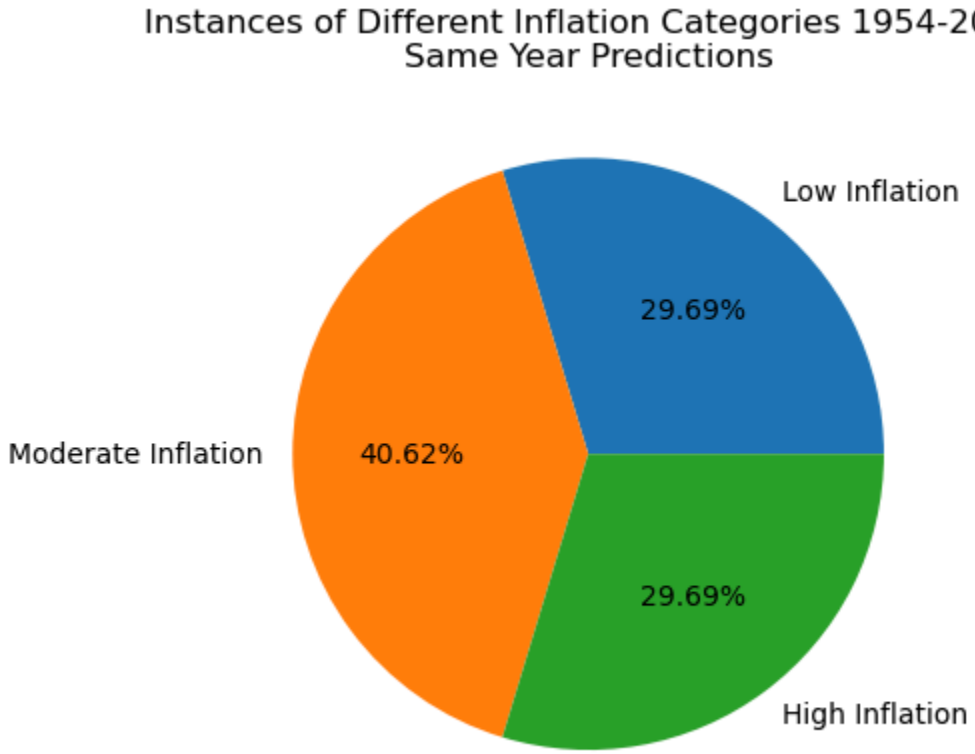
As seen, the MSE (mean squared error) went down after including GDP, and the R2 (variance) went up, indicating that GDP is actually a predictive factor for future inflation, as it explains more of the variance and there is less overall error in the models predictions, however the effect is small, and interest rates seem to be by far the most predictive factor (out of these two features). This makes sense, as you would expect a generally lower inflation after high gdp years, and higher inflation after generally low gdp years (cooling vs heating up the economy), but since U.S. GDP has been on such a smooth uptrend, the effect is very minimal. Here is a visual aid to make better sense of the U.S. GDP rise over time:



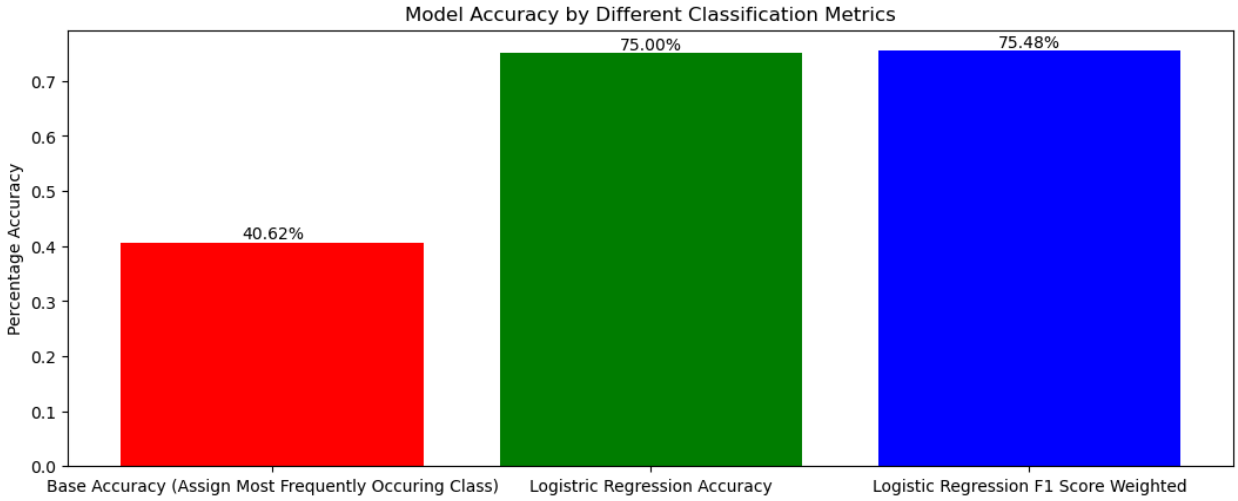
Methods:

After significant exploration of the data we finally decided to train a model using the logistic regression classifier (a supervised learning algorithm) in order to make predictions for inflation categories (low <2%, moderate 2-4%, and high >4%) since it was the most logical extension to our linear regression analysis during the data exploration phase. Initially had decided to make predictions for inflation of the current year based on the current years data (a mistake on our part, since we were trying to predict future inflation, i.e. the next year, but this ended up proving useful to see how model accuracy diminishes with each passing year). Here are some visual aids to illustrate the original (yet flawed) methodologies success:

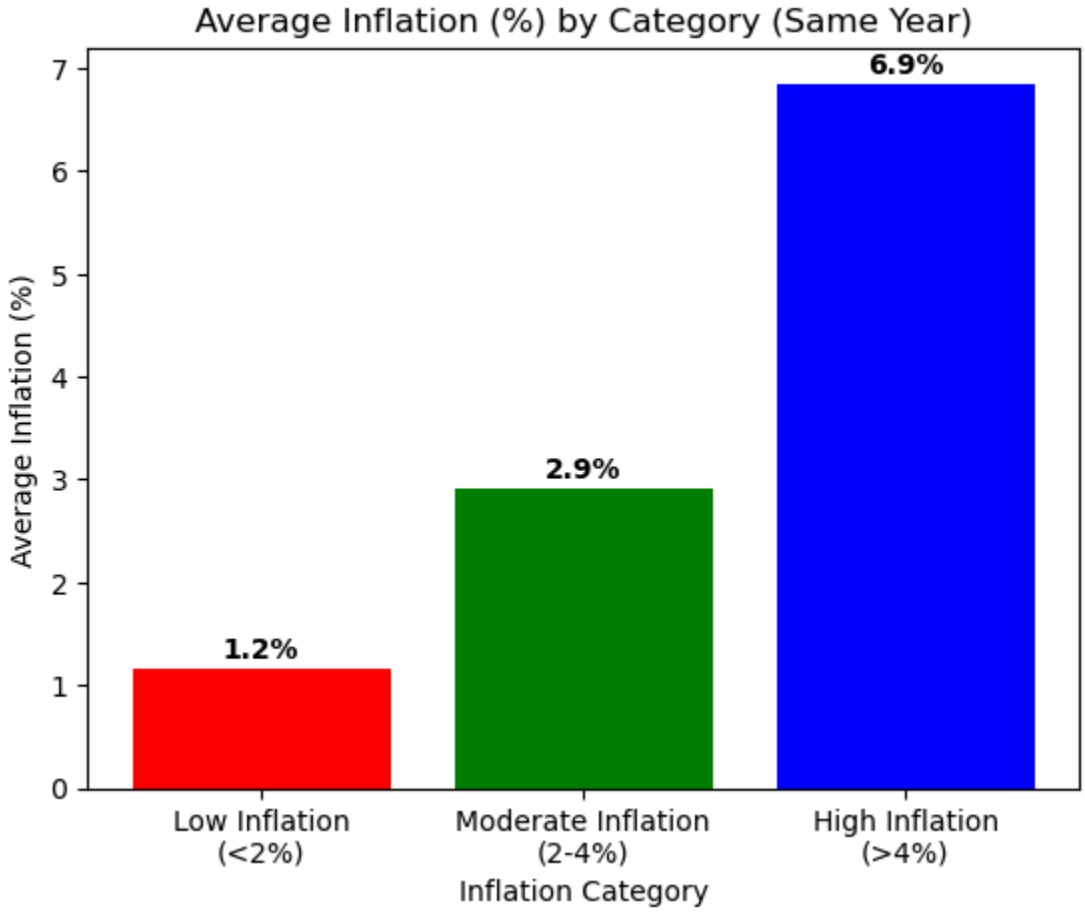
Distribution of Inflation Categories (Same Year Prediction):



Validation metrics (Same Year Prediction):

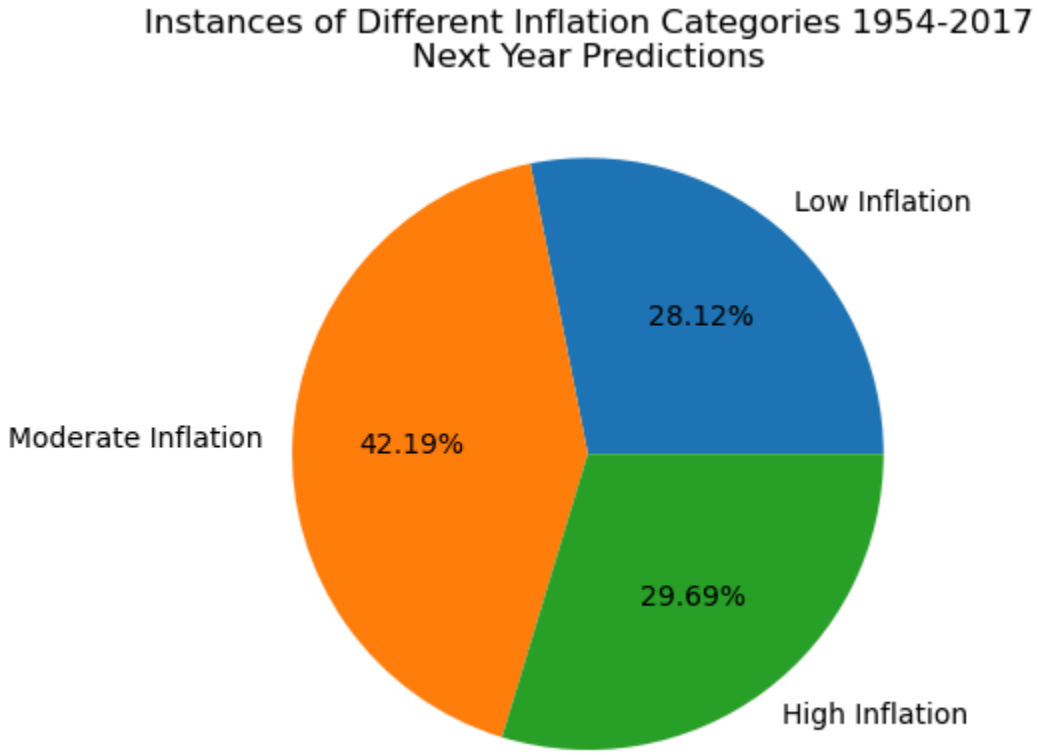


Average Inflation Percentage by Category (Same Year Prediction):

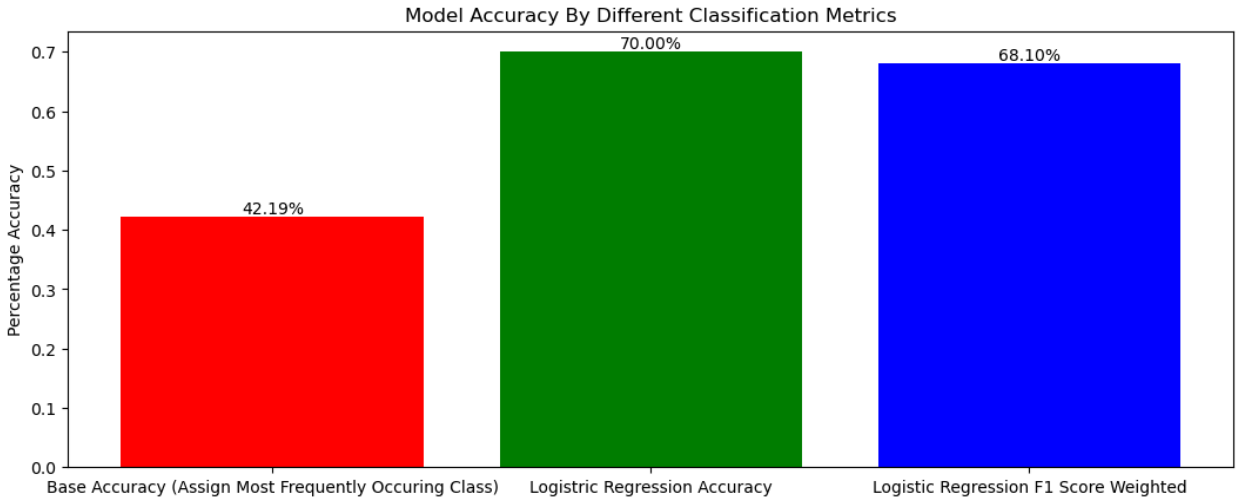


Now when actually looking at predicting the next years inflation based on the current years inflation data, the validation metrics worsened (but not by much), as can be seen here:

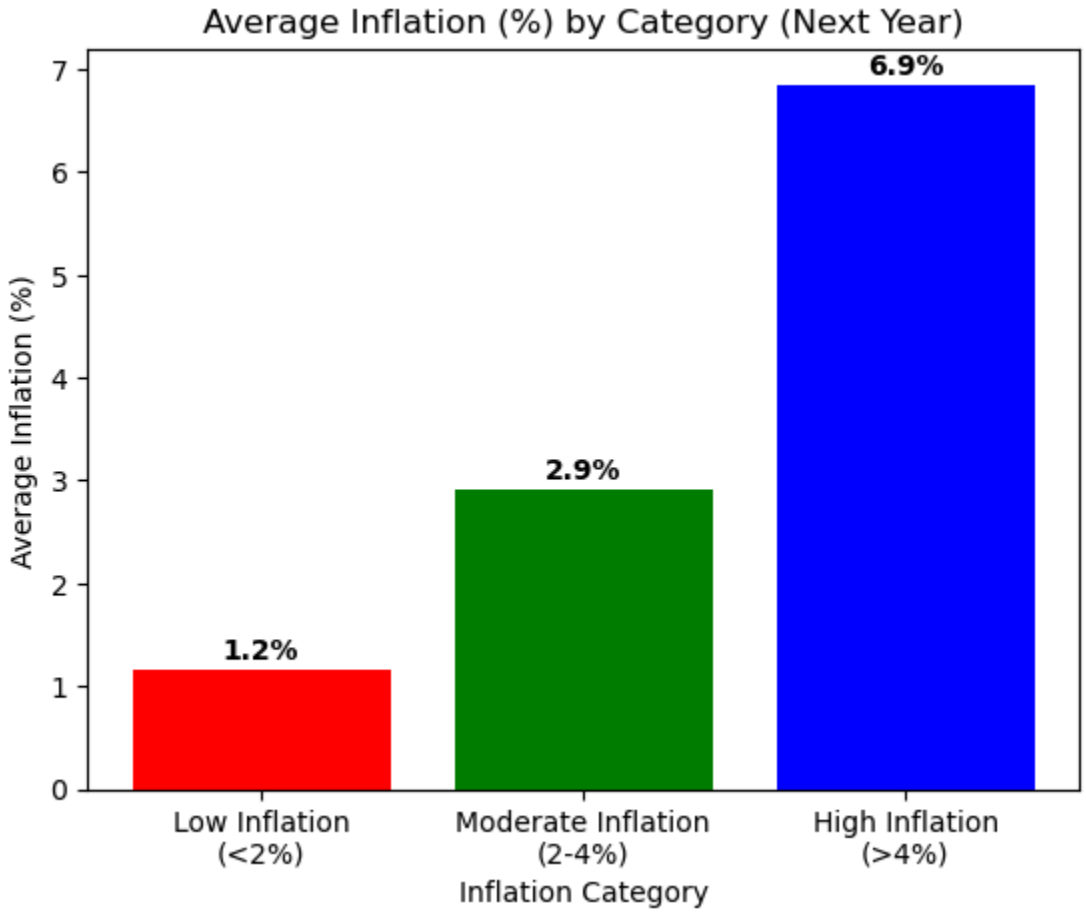
Distribution of Inflation Categories (Next Year Prediction):



Validation metrics (Next Year Prediction):



Average Inflation Percentage by Category (Next Year Prediction):



As can be seen, the logistic regression did a good job of predicting what category inflation would fall into (low <2%, moderate 2-4%, or high >4%) as validated by both accuracy and f1-score relative to the base accuracy, accuracy boosts of 27.81% for accuracy 25.91% for f1-score respectively, which takes the model from being no better than a coin toss (in the case of just picking the most common category) to being able to base decisions off of (using the logistic regression model). This is obviously true when taking into account the fact that this second run through is looking at next year inflation predictions, not the current year as was done at first (flawed), providing legitimately useful information to any consumers.

Analysis of Results:

Interest Rate Only Analysis: Our first predictive experiment employed a simple linear regression model using only the Effective Federal Funds Rate to forecast the next-year inflation. On the held-out test set, this interest-rate-only model yielded a MSE (Mean Squared Error) of 3.029 and an R2 of 0.594, indicating that a little over half of the year-to-year variation in inflation could be explained by changes in the federal funds rate alone. While these results confirmed a statistically significant relationship between interest rates and inflation, they also underscored that much of the year-over-year fluctuation remained unaccounted for by a single feature.

Interest Rate + GDP Analysis: To test whether additional macroeconomic context would improve predictive power, we next augmented the model with current-year GDP growth as a second feature. Incorporating GDP led to a barely noticeable (but present) performance boost. MSE fell to 2.991, a 1.27% reduction in mean squared error, and R2 climbed to a 0.599. This still means a little over half of inflation variability was being captured by our two-feature regression, albeit with slightly more of the variance being explained when taking GDP into account as well. The simultaneous decrease in error and increase in explained variance demonstrated that GDP growth, while a subler driver than interest rates, contained complementary information that enhanced our inflation forecasts.

Validation + Performance Review: Our goal was not raw point-prediction, but rather categorizing next-year inflation into actionable bands (low <2%, moderate 2–4%, high >4%), so we recast the problem as one of classification using logistic regression. Using the same annual rate and GDP input, a baseline classifier that always predicts “low inflation” achieved only 42.19% accuracy. To compare, our tuned logistic regression model delivered 70.0% overall accuracy and an F1 score of 68.1% on the test set. This represents a 65.91% relative improvement over the baseline (in terms of accuracy), and 61.41% improvement over the baseline (in terms of F1 score). This improvement transforms predictions from “worse than a coin flip” to a model that correctly classifies nearly seven out of ten years.

Validation and Testing:

We used both MSE (Mean Squared Error) and R-Squared metrics in order to test the accuracy and validity of our linear regression model when using both interest rates and GDP as our inputs for predictive factors. R-Squared was used for assessing variance and explaining the influence of our inputs on the model, while MSE was used for assessing accuracy and detecting the presence of outliers. Comparisons can also be drawn between the different results of training and testing.

R-Squared Validation:

R-Squared is the more important metric here since it can be used to explain what percentage of inflation is controlled by the interest rates or GDP. When using only interest rates, the R2 value in testing ended up being 0.594702, which indicates that the interest rate had a

noticeable impact on inflation rates. When also using GDP, the result mostly stays the same, slightly raising to a value of 0.599753. As previously explained in the Exploratory Analysis of the Data section, the impact of adding the GDP as a predictive factor is minimal, indicating that the interest rates are the dominant factor in impacting inflation. When comparing the training R2 and test R2 results for both the interest rates and GDP, we found that both cases had a slight decrease in the result when going from training to testing, with the interest rates decreasing from 0.603937 in training to 0.594702 in testing, and the latter decreasing from 0.603997 in training to 0.599753 in testing.

Mean Squared Error Validation:

MSE was used to measure the quality of our model's output, being chosen since it heavily penalizes outliers, making them generally easier to detect. The MSE when testing interest rates as a predictive factor ended up being a 3.029354, while the test MSE for using GDP as well ended up being a 2.991599. This means that although there is a presence of errors in our output, it only had a minimal impact on the results, meaning the model could still draw accurate predictions. Comparing the training and test MSE results for both the interest rates and GDP shows very similar results, with the MSE values for interest rates increasing from 2.927423 to 3.029354, and the MSE values for GDP increasing from 2.926975 to 2.991599. Ultimately, the MSE ended up not being as important of a factor to the model's output when compared to R-Squared.

Discussion and Conclusions:

With all of the data we collected, it is fair to say that people could benefit from using this model. We have found that there are some deciding factors when it comes to predicting the inflation rate, based mainly on interest rates and GDP. Due to this, it can be deduced to a certain degree when inflation can have a more major effect on the economy. This can save many Americans the headache of trying to find when the economy may inflate in a more prevalent manner. There are many ways people can benefit from this information. They could benefit by saving money on certain products, such as (fill in product and savings), because of the accuracy of the relationship between interest rates and inflation, you can find out where you can save on products. They can benefit from buying major products such as houses and property since inflation has a major effect on the housing market. Buying things from outside the US is another area to look at, since many things can be bought for either cheaper or more expensive.

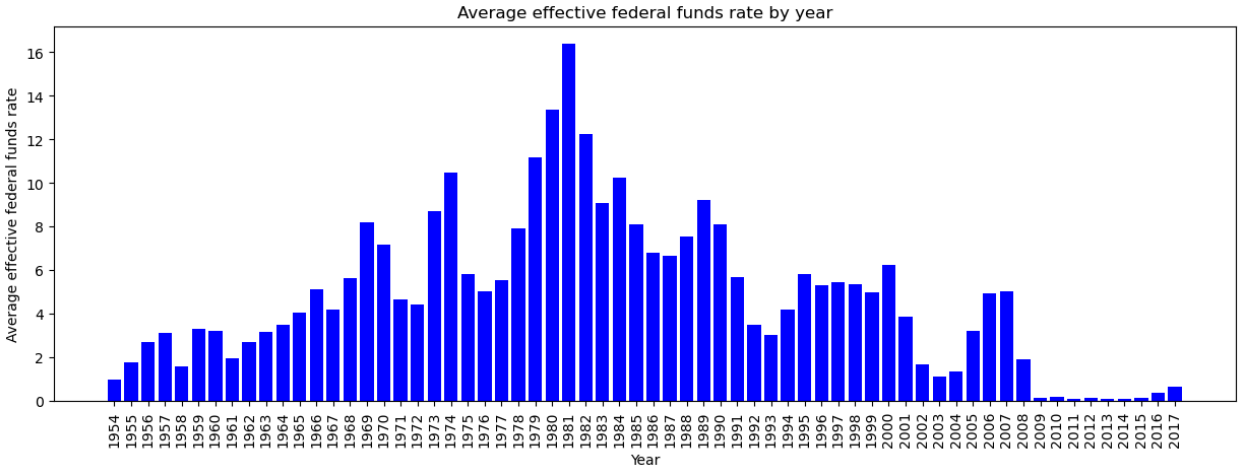
To conclude this report, the data and information found throughout this project can be used to somewhat accurately tell if inflation is going to increase or decrease based on the two main deciding factors, GDP and interest rates. The best case shows that when interest rates increase or decrease, so too does inflation to some degree. On top of that, we have also found that GDP can affect these rates as well, since the more money the US is bringing in, the less the federal rates will need to be increased for that specific year, thus bringing the inflation down

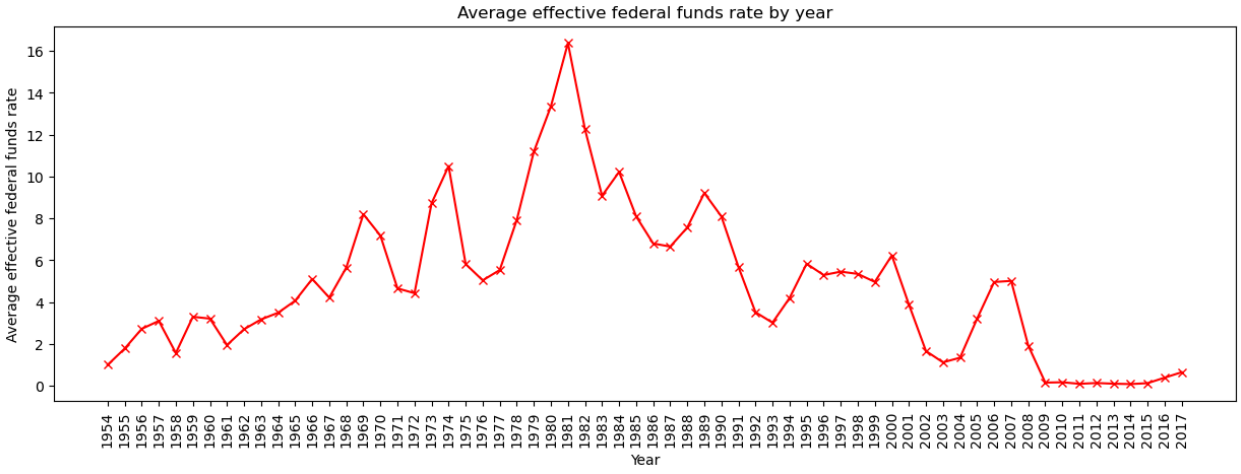
slightly as well. This is important information for many due to the ramifications that this can hold over the average citizen attempting to understand something as complex as this.

Appendix:

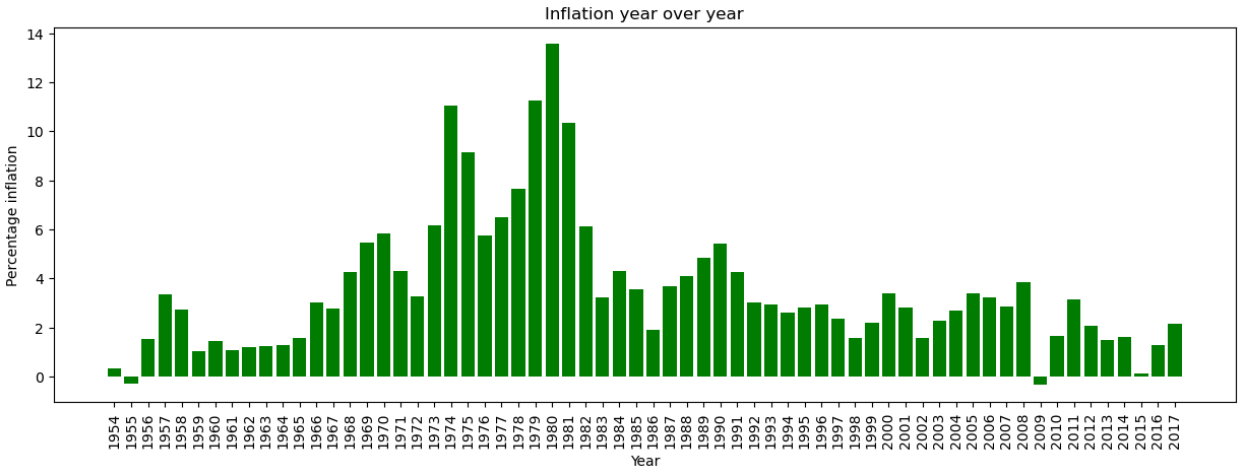
Images for Visual Aid:

Average interest rates from 1954-2017:

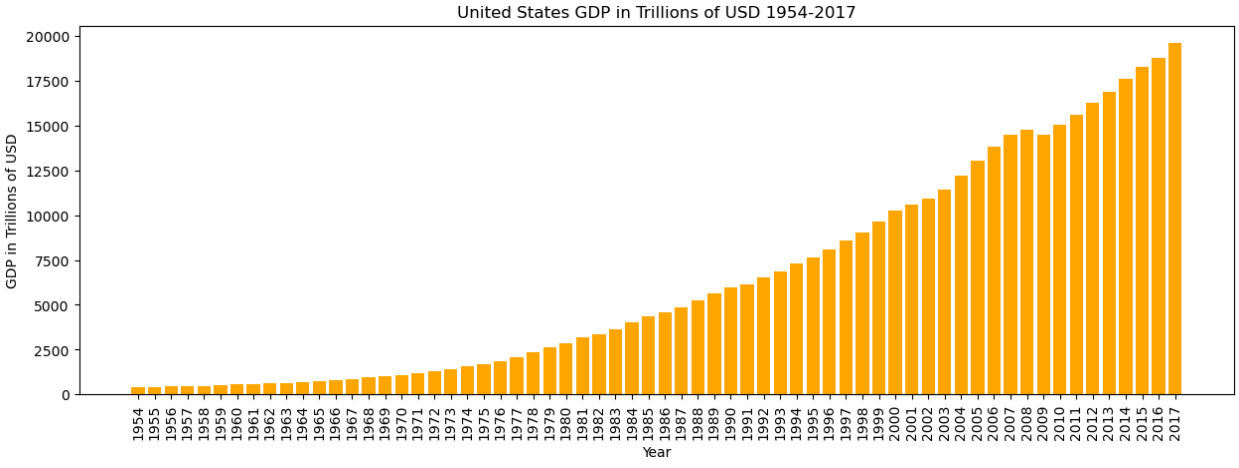


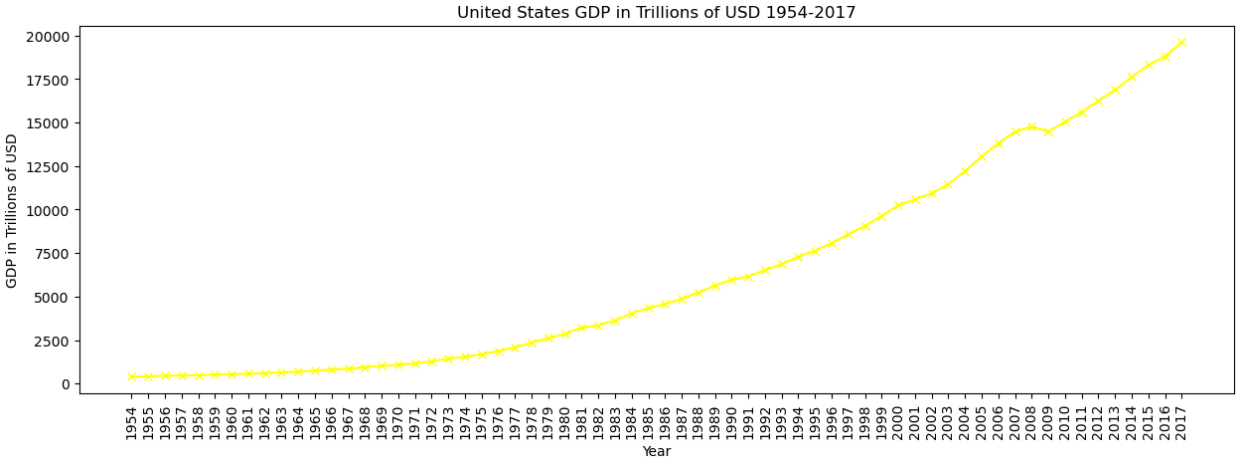


YoY Percentage Change in Inflation:

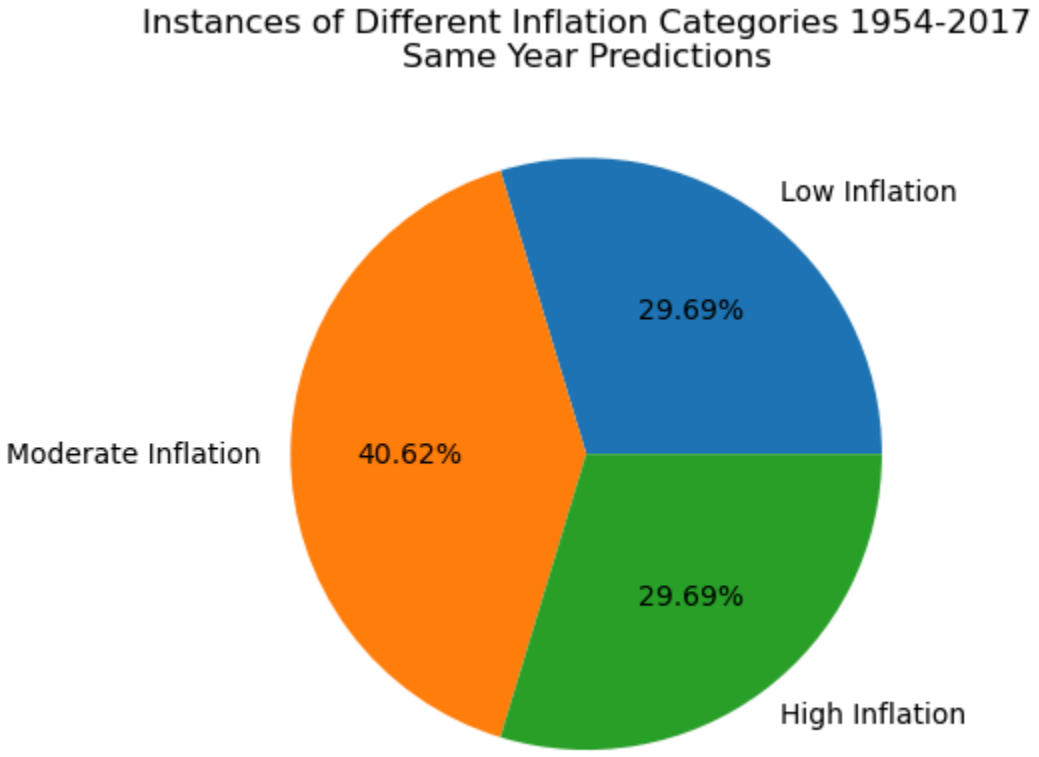


United States GDP in Trillions of USD 1954-2017:

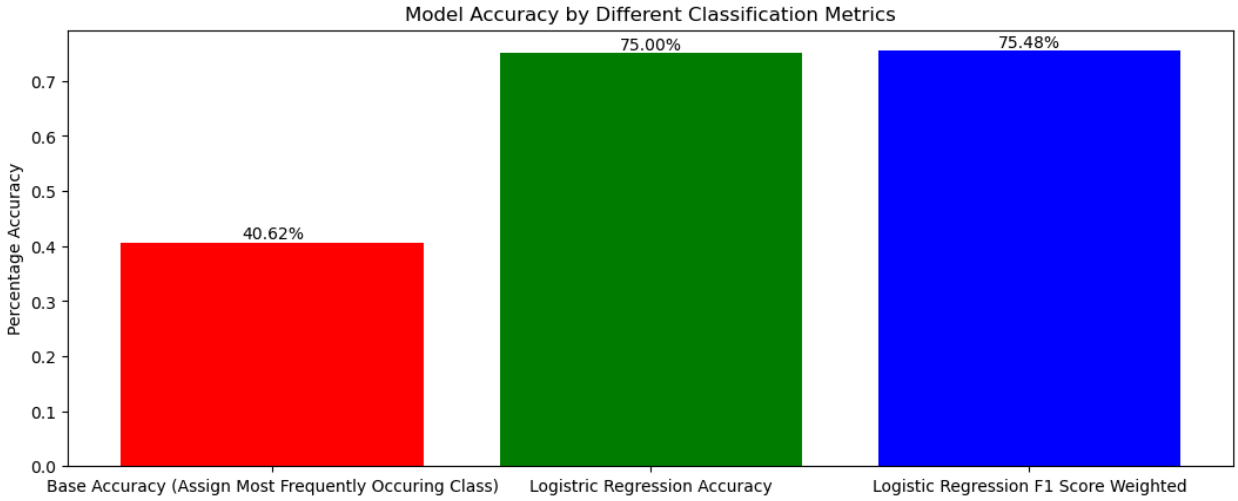




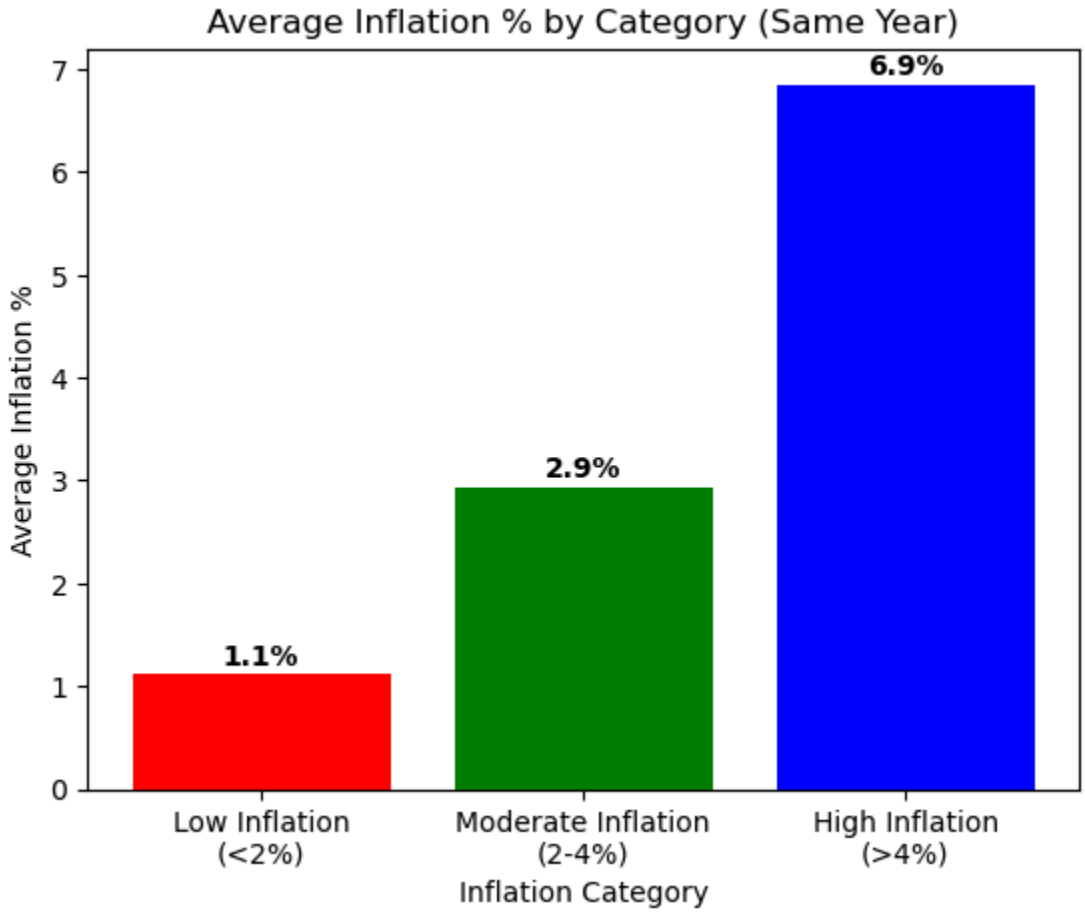
Distribution of Inflation Categories When Doing Same Year Predictions:



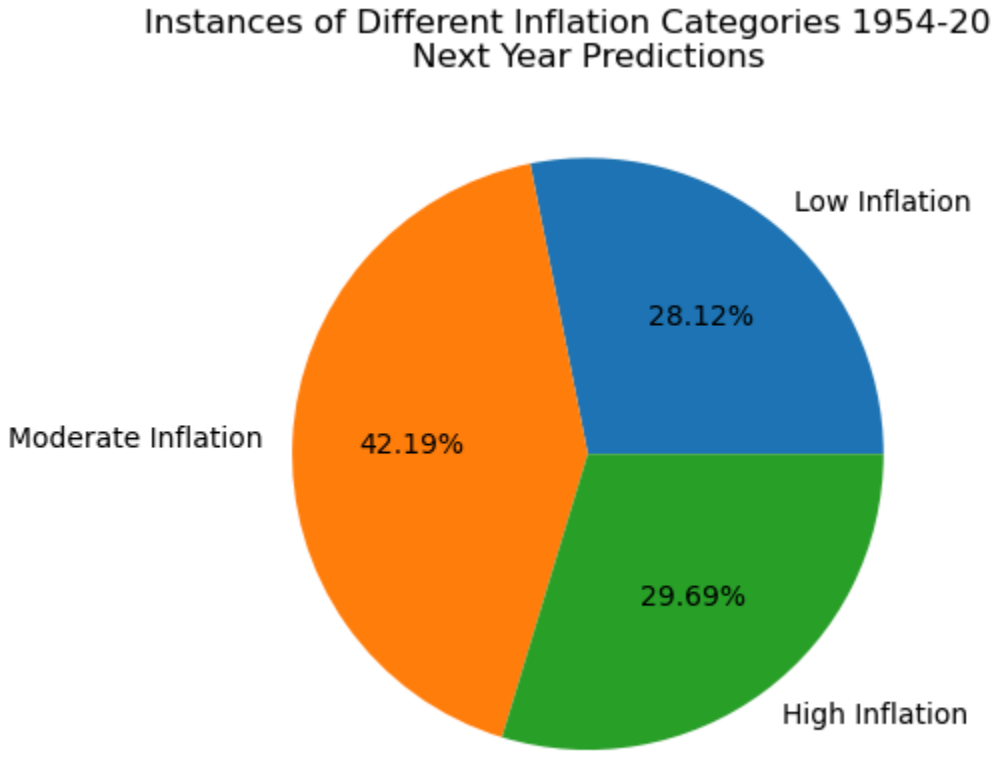
Model Accuracy by Different Validation Metrics Doing Same Year Predictions (Accuracy + F1 Score):



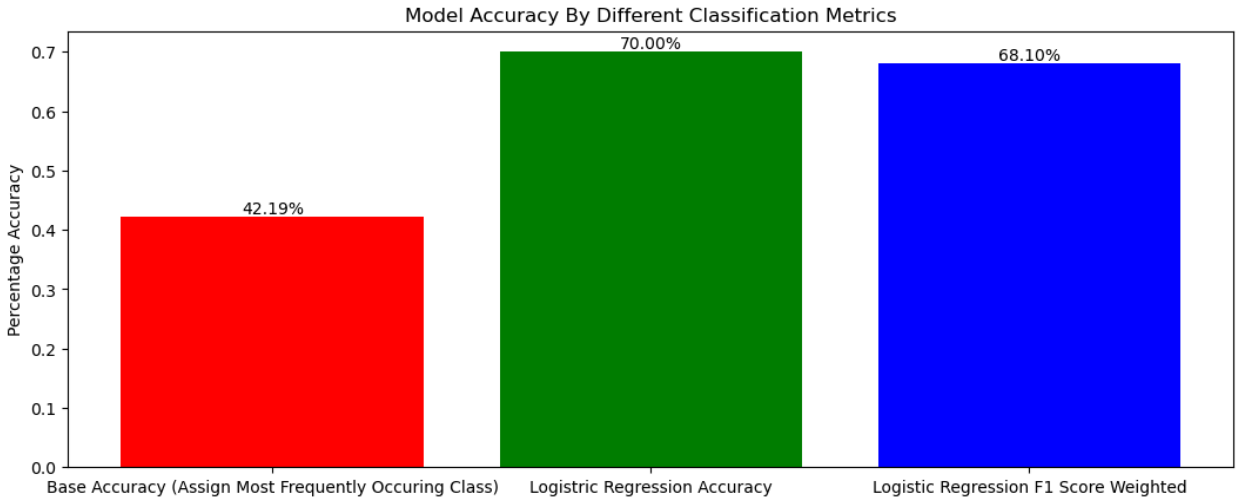
Average Inflation by Category Doing Same Year Predictions:



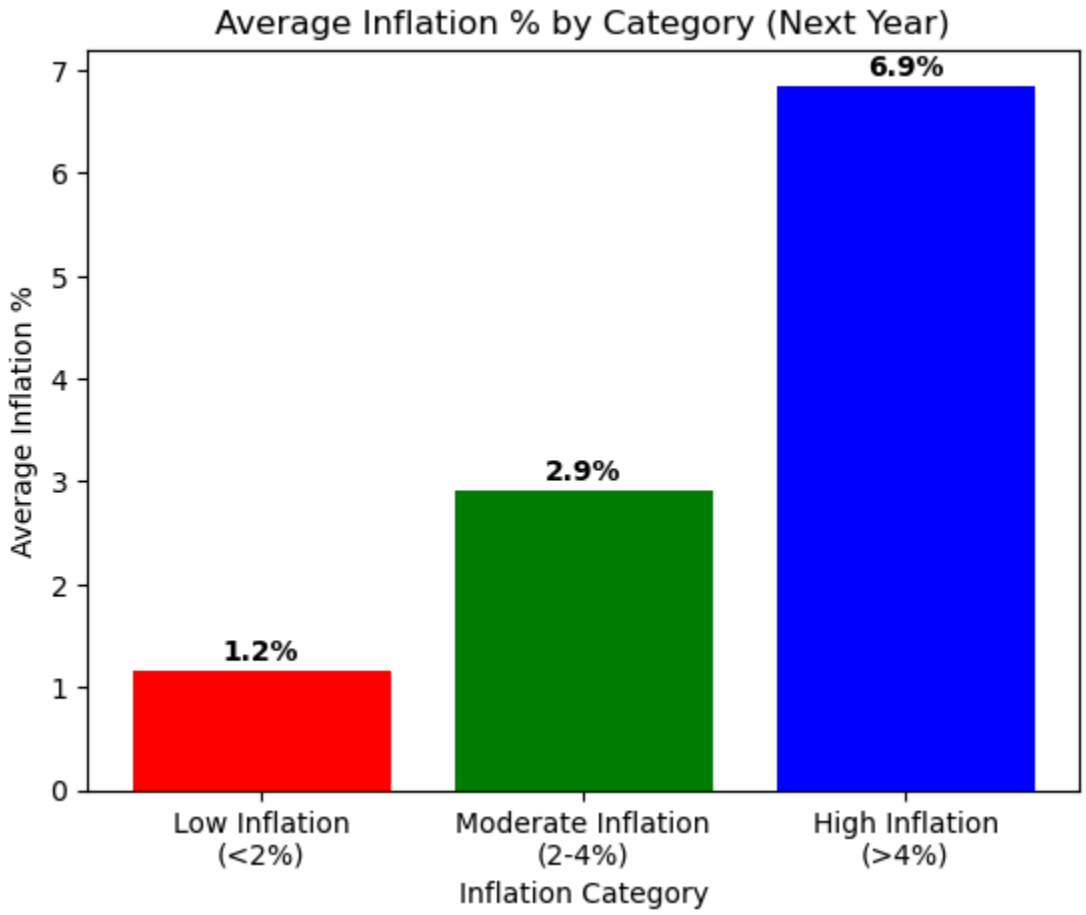
Distribution of Inflation Categories When Doing Next Year Predictions:



Model Accuracy by Different Validation Metrics Doing Next Year Predictions (Accuracy + F1 Score):



Average Inflation by Category Doing Next Year Predictions:



Sources:

1. Inflation Dataset 1913-2022 (Kaggle): https://www.kaggle.com/datasets/neelgajare/usa-cpi-inflation-from-19132022
2. Interest Rate Dataset 1954-2017 (Federal Reserve): https://www.kaggle.com/datasets/federalreserve/interest-rates
3. USA YoY Inflation + Federal Reserve Interest Rates 1954-2017 (Kaggle): https://www.kaggle.com/datasets/pumpkinsmith/interest-rate-yoy-inflation-data-1954-to-2017
4. USA Historical Income Tax Rate Data 1913-2020 (Kaggle): https://www.kaggle.com/datasets/frtgnn/historical-income-tax-rates-brackets
5. USA GDP Data 1947-2025 (Federal Reserve): https://fred.stlouisfed.org/series/GDP
6. USA CPI Data 1947-2025 (Federal Reserve): https://fred.stlouisfed.org/series/CPIAUCSL
7. USA Federal Reserve Interest Rates 1954-2024 (Federal Reserve): https://fred.stlouisfed.org/series/FEDFUNDS