UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística (FME)
Màster en Estadística i Investigació Operativa

Programació i Bases de Dades Estadístiques

Final Project
Outcome Report

Autors: Leonel Fernando Nabaza Ruibal

Professor: Dr. Àlex Barceló Cuerda

Barcelona, December 26, 2025

Resum

Aquest projecte analitza l'impacte de diversos indicadors socioeconòmics en l'atur als EUA durant un període de 25 anys (2000–2025), utilitzant dades de la Reserva Federal i Yahoo Finance. S'ha emprat un Random Forest Regressor per predir les tendències de l'atur i identificar-ne els factors clau, revelant que els tipus d'interès i el sentiment del consumidor són els elements més crítics. A més, s'ha dut a terme un model lineal i un test d'hipòtesi formal, verificant estadísticament que l'índex SP500 és una variable significativa per modelitzar la dinàmica de l'atur.

Abstract

This project analyses the impact of socioeconomic indicators on US unemployment over a 25-year period (2000–2025) using data from the Federal Reserve and Yahoo Finance. A Random Forest Regressor was employed to predict unemployment trends and identify key drivers, revealing that interest rates and consumer sentiment are the most critical factors. Furthermore, a linear model and a formal hypothesis test were conducted, statistically verifying that the SP500 index is a significant variable for modelling unemployment dynamics.

Contents

Chapter 1

Introduction

The aim of this project is to first study the effect that several socioeconomic data have over the unemployment in the United States. To tackle this challenge, a Random Forest Regressor will be trained under data coming the Federal Reserve of Economic Data and from Yahoo Finance. Since we are working with non synthetic data, doing a relatively small selection of features and trying to predict the unemployment that there'll be next, the assessment of the results will not be very strict on having very low errors, but in correctly guessing the tendency of the evolution of unemployment.

The second aim of the project is to build a linear model for the same purpose, but with the intention of statistically studying if including the variable related to the price of SP500 bonds is necessary or not. For that, knowledge from the subject Linear and Generalized Linear Models will be used. So, a hypothesis test will be performed on the coefficient of the variable.

Chapter 2

Data for the model

The selected data comes from the two following sources: yfinance, which is is a popular, free, open-source Python library that lets developers easily download historical market data, financial statements, and other relevant info directly from Yahoo Finance for analysis, research, or building trading strategies; and pandas-datareader which is an up-to-date remote data access for pandas, and among its sources, there was data from the Federal Reserve Economic Data.

As mentioned in the introduction, I decided to study the target variable unemployment, which came from the pandas_datareader package with respect to the variables:

- CPI_Inflation, from pandas_datareader. It is defined as the rate at which the average price of a fixed 'basket' of consumer goods and services (food, housing, transport, healthcare, etc.) rises or falls over time.

- Initial_Claims, from pandas_datareader. It represents initial claims for unemployment insurance in the United States, adjusted for seasonality.

- Interest_Rate, from pandas_datareader. It refers to the interest rate at which U.S. banks and financial institutions lend reserve balances to each other overnight. It is set as a target range by the Federal Open Market Committee.

- Consumer_Sentiment, from pandas_datareader. It is an indicator that measures how optimistic or pessimistic households are about the economy and their own financial situation. It is typically measured through survey-based rather than observed transactions.

- Industrial_Production, from pandas_datareader. The Industrial Production Index is a macroeconomical index that measures the real level of production of the industrial sector of the US. It includes manufacturing, mining and utilities (electricity, gas or water).

- `SP500_Index`, from `yfinance`. Note that several columns can be obtained when importing this dataset. Because it is the most informative and to avoid redundancy, I selected the closing price of these bonds.

All these data is only from the US.

I selected a period of time of 25 years, from the 1st of January of 2000 till the 21st of December of 2025. This ensures a wide range of data and values, attending that several crisis and economic periods happened during that time (the real estate bubble or the pandemic, for example).

Because of the different periodicities among the features and the target, the data was outer-joined and then resampled so that everything was on a monthly format. Moreover, a one month shift was applied to the target variable, since our interest is to predict the unemployment next month.

Chapter 3

Random Forest Regressor

I decided to use a Random Forest Regressor because it can capture non-linear relationships and complex interactions between the variables and the target without imposing strong parametric assumptions. It is robust to multicollinearity and noise, handles mixed-frequency economic indicators well after aggregation, and typically delivers strong out-of -sample performance, making it suitable for predictive for this task.

The metrics for testing have been the MAPE (Mean Absolute Percentage Error) and MAE (Mean Absolute Error). These performance assessors are complementary to understand the results. In this case, I obtained a MAE of 1.15 (which means that the average error in predicting the percentage of unemployment is of 1.15%) and the MAPE of 27.03%.

The obtained results are positive taking into account that the set of features is reduced (I could have chosen many more), have different periodicities and some of this data is hard to measure accurately, because it may come from social studies or because it compromises many parts (such as many companies). Furthermore, one the aspects I was interested in, is that the model correctly guess the tendency of unemployment (if it would go up or down), and it work well in that regard, as it can be appreciated in the following figure:

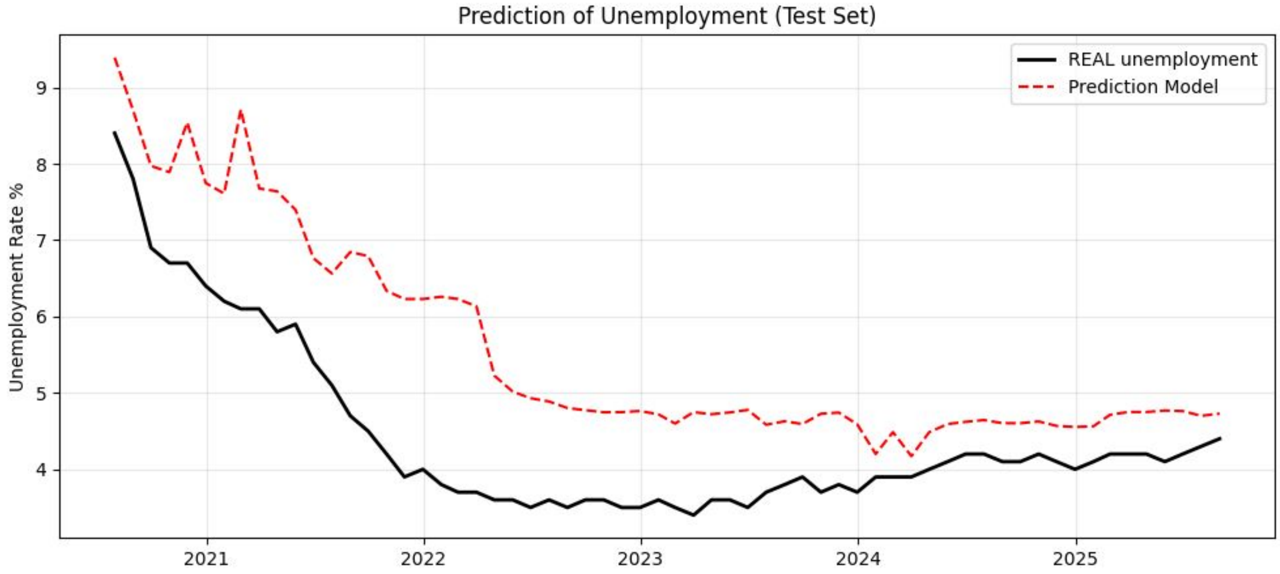

Figure 3.1: Result of predicting the unemployment with the Random Forest Regressor

Finally, by printing the importance of the features for the model, I could distill that the most essential ones for making predictions are the Interest rate, the Consumer Sentiment and the Initial Claims, while Industrial production, the SP500 bonds and the Inflation are considerably less significant. This can be seen in the following plot:

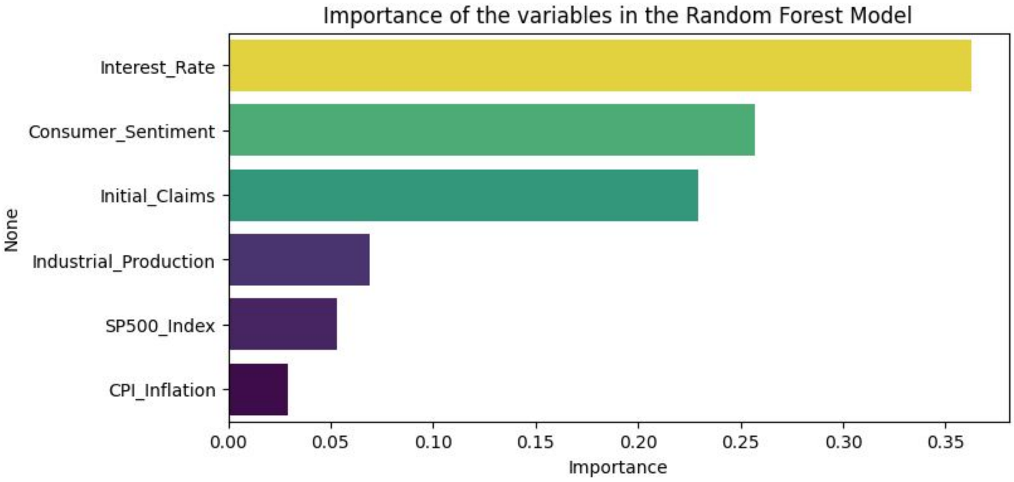

Figure 3.2: Importance of the features

Additionally, because of the reasons explained in the journal, a Random Forest Regressor was also implemented for predicting unemployment the same month, having much lower errors (MAE of 0.7590 and MAPE of 17.76%):

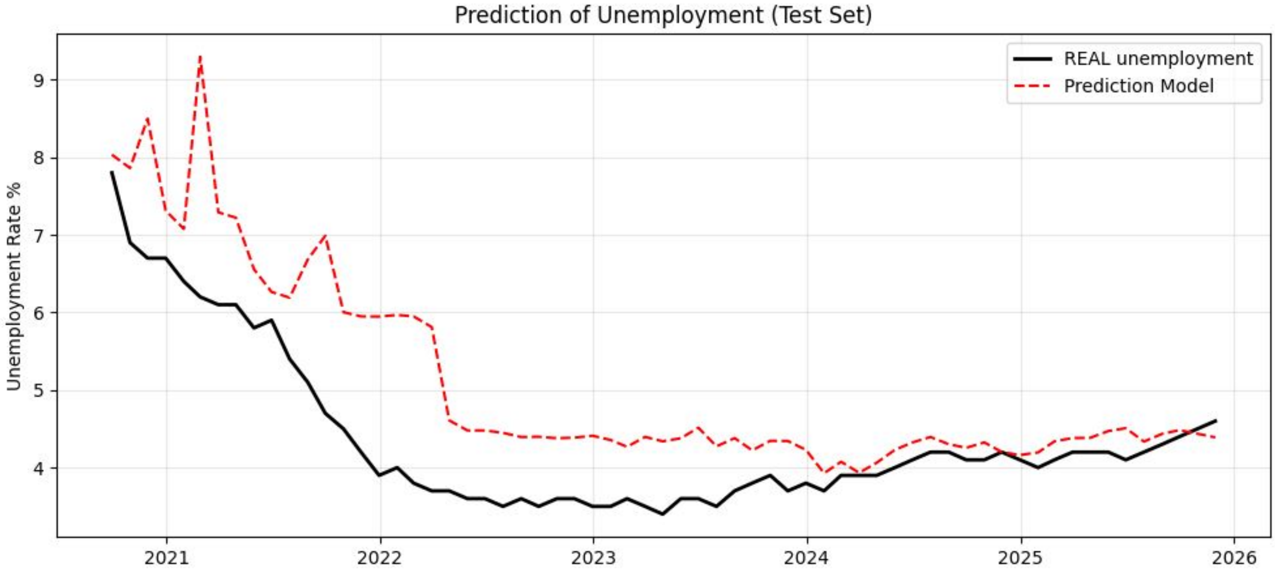

Figure 3.3: Random Forest Regressor applied to predict same month unemployment

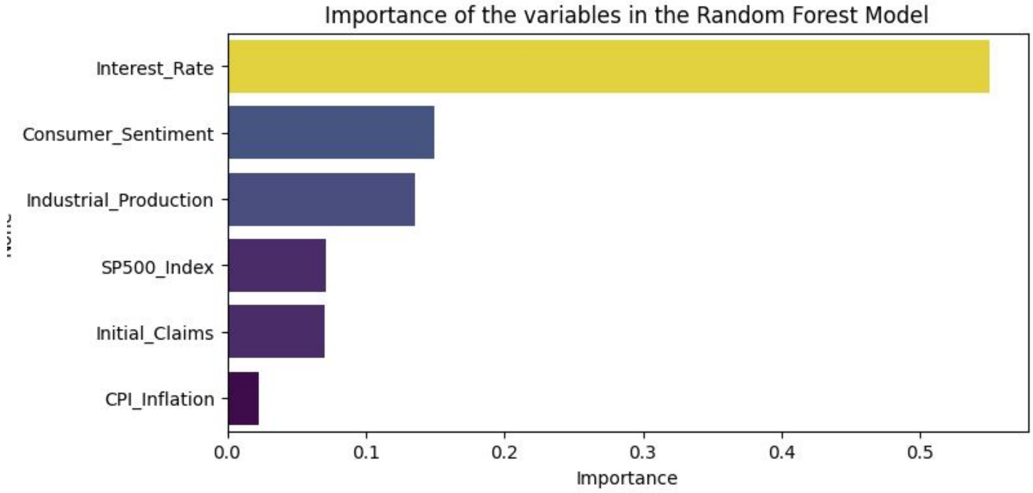

Figure 3.4: Importance of the features with no shift in unemployment

The main additional observation (besides the results) is about the importance of the Interest Rates, for which the following conclusion was reached: higher rates increase borrowing costs, slowing business investment, consumer spending, and growth, leading to reduced hiring or layoffs, while lower rates decrease borrowing costs, encouraging investment, boosting demand, and spurring job creation, thereby lowering unemployment.

Chapter 4

Hypothesis test

In this part I built a linear model with the same dataset and studied if the coefficient for the variable `SP500` should be zero or different from zero; i.e., if that variable should be included or not in the model. For this purpose I followed the following mathematical framework (assuming this time that the data follows a linear model distribution).

We have a set of observations contained in X, where each column contains a feature (and the first column is full of ones, because it is the one for the intercept) and we have p features, and each row is an observation (a particular set of values of the features), with n rows in total. We want to predict a target variable y, that takes value y_i for each observation. We are building a model under the assumption that any y_i can be expressed as a linear combination of our features: $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (a linear model).

Now, imagine we have built that model and, by the method of least squares, have obtained the best possible set of coefficients $\hat{\beta}$ for our linear model. The study case we are doing here is checking whether if for β_j it should be better to make it zero (exclude the feature from the model) or not.

Recall that $T \sim t_\nu$ (Student t with ν degrees of freedom) is defined as follows:

$$T = \frac{Z}{\sqrt{Q/\nu}}, \qquad Z \sim \mathcal{N}(0,1),\ Q \sim \chi^2_\nu,\ Z \text{ independent of } Q.$$

We have the following distributions for $\hat{\beta}_j$ and $\hat{\sigma}^2$ (the estimator of the variance):

$$\hat{\beta}_j \sim \mathcal{N}\big(\beta_j,\ \sigma^2 (X'X)^{-1}_{j+1,j+1}\big)\,, \qquad \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1},$$

where $\hat{\beta}_j$ and $\hat{\sigma}^2$ are independent.

Standardizing $\hat{\beta}_j$, we find

$$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 (X'X)^{-1}_{j+1,j+1}}} \sim \mathcal{N}(0,1).$$

Now, define $Q = (n-p-1)\hat{\sigma}^2/\sigma^2$ and write

$$T_j = \frac{Z}{\sqrt{Q/(n-p-1)}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{j+1,j+1}}} \sim t_{n-p-1},$$

from which we can construct hypothesis tests and confidence intervals for β_j.

If we want to test $H_0 : \beta_j = b_j$ against $H_1 : \beta_j \neq b_j$ for a known constant b_j at a significance level α, the decision rule for rejecting H_0 is

$$\text{reject } H_0 \iff |T_j| \geq q_t(1-\alpha/2,\, n-p-1).$$

The procedure of training and testing was exactly the same as the one for the previous model, but selecting a linear one instead. The obtained coefficients for the model can be seen in the following figure, and we can observe that there is a change that our study variable is not important for it, since its value is very low:

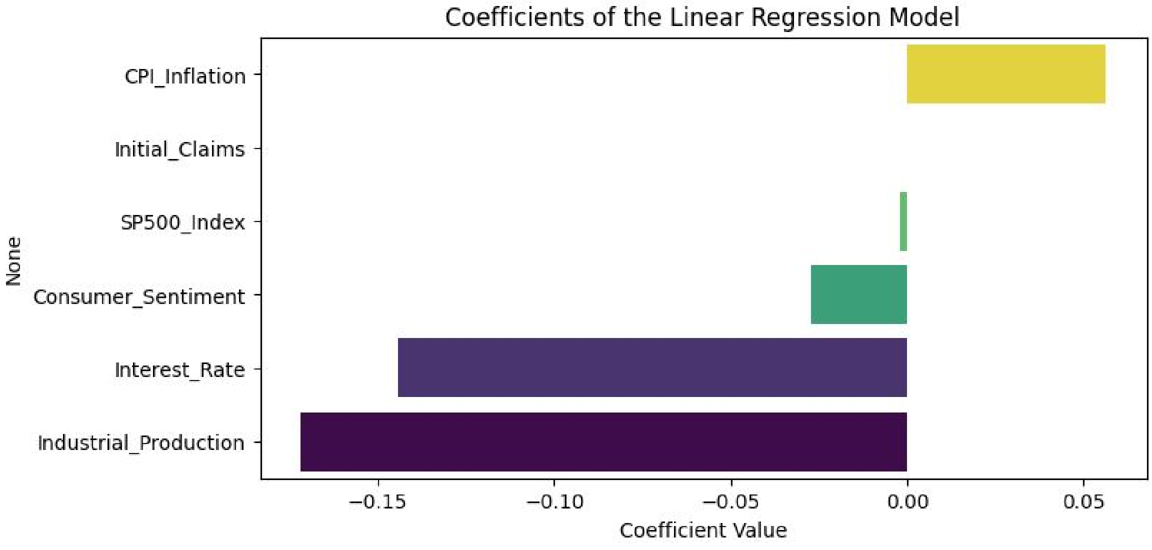

Figure 4.1: Coefficients for the linear model

Nevertheless, the result of my test was that variable `SP500` should not be excluded from the model, since the value of T_j was -6.5177 (much lower than the left boundary, -1.9699). Moreover, to further reinforce the results. I created a plot that compares the evolution of

`Unemployment_Rate` and `SP500` over time, and it can be clearly seen that there exist some negative dependency (when one variable goes down the other one goes up, and viceversa).

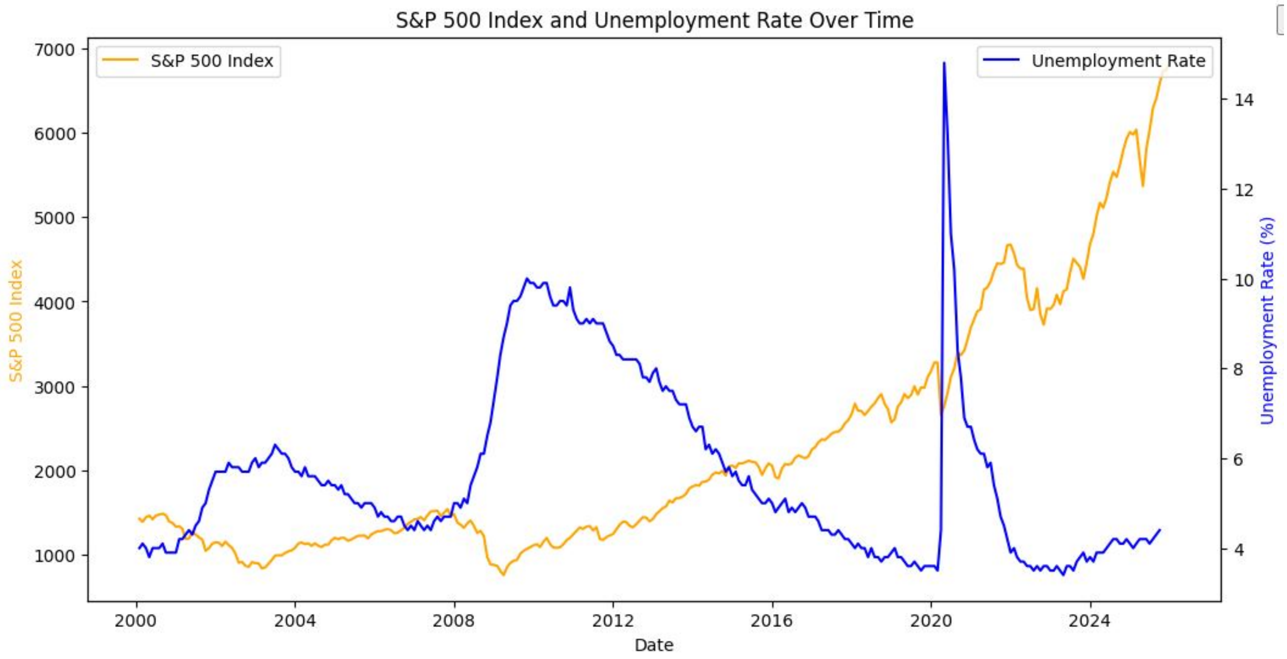

Figure 4.2: Evolution over time Unemployment and SP500

Chapter 5

Conclusions

The project successfully accomplished its primary objectives of analysing the impact of socioeconomic data on US unemployment through a workflow involving data aggregation, machine learning, and statistical inference. The work involved collecting 25 years of indicators—such as Interest Rates, CPI, and the SP500—from sources like the Federal Reserve and Yahoo Finance. A Random Forest Regressor was then trained to predict future unemployment, successfully capturing the trend evolution with a Mean Absolute Error of 1.15%. This analysis identified that Interest Rates and Consumer Sentiment are the most critical features for predicting unemployment fluctuations.

Furthermore, the project fulfilled its secondary goal of statistically validating the relationship between financial markets and labour statistics by constructing a linear model to perform a formal hypothesis test. The investigation specifically focused on the SP500 index variable, yielding a statistic of $T_j = -6.5177$, which necessitated the rejection of the null hypothesis. This result confirmed that the SP500 is a statistically significant variable that should be included in the model, reinforcing the visual evidence of a clear negative dependency between stock market performance and unemployment rates.

Chapter 6

Appendix

For the corresponding project: see https://github.com/LeonelFNR/PBDE-Project.