Project 2

With recent news spiraling through multiple funnels in the media, the topic of immigration is a hot button topic that dominates not only the media, but economic, political, and especially individuals personal gain. Let it be moving to a new country for a better education, safety, or gain a new experience.

This intern leads to the topic of immigration, which asks a number of questions from: Why should we let this individual in? How will they affect the economy? and a number of other questions.

I have specifically targeted predicting the US immigration quota for next 10 years. This project has a special place in my heart because of these two. My dad immigrated here in the 70s, and my mom did in the 90s. Without them, who knows if I would’ve found out the novel idea of not only what a taco is, but putting breakfast in that taco and calling it a breakfast taco.

With that background, I’m Cyrus Rustomji, and I’ll go into a little more detail to how I was able to predict the US immigration quota in the next 5 years.

I used the past 50+ years of data, and I chose US GDP, US population, global GDP, global population, and a foreign policy indicator. Both GDPs are adjusted for inflation, and the foreign policy indicator is binary meaning a 0 if president has a strict immigration policy and a 1 if the president has a lax immigration. As I cut the years in my model down to bring a more modern approach to the immigration policy, 50% of presidents in my model have a strict policy while 50% have a lax one.

When I first plugged in the inputs for my model which is the OLS and Linear regression, both my R^2’s were high, which are both a measure of how statistically significant my model is. The closer to 1 then better. This wasn’t enough though, so I did a train/test split of the data. I trained 80% of the data, and after building a model from the trained data, I was able to use the test data to validate my model to make sure there were no discrepancies. Then I used a regularization, which is a method to bring the coefficients closer to zero. I used an Elastic Net when I sized which is a form of regularization, and I’ll get to later show you how this Elastic Net came into play.

The graph on the right shows how the US population has changed over the past 10 years. You can tell that it doesn’t change completely change from light blue to dark blue. This takes into consideration how the GDP did not linearly during 2008 due to the Recession.

This slide shows my coefficients using Elastic Net. They went from \_\_\_\_ to extremely close to zero. This was able to bring my accuracy score to a 98% from a 87% when I just used linear regression. The foreign policy indicator seems to be high, but keep in mind that it’s binary and the data is a 50/50 split, so it’s coefficient does not skew the model.

Ways to improve the model, I think if I have refugee data and H-1B data for the past 50 years, and each being specified by country; I’ll be able to better show how and why individuals come into the US.

Sources I used a government site, and a data aggregation site for to pull all the numbers for this project.

So no matter who the president is, where the GDP may lie, and plus with a constant rise in population, we should consistently see a rise in immigration every year, and thus boosting the economy, our political power, and individuals personal gain.

With that, any questions?

* Talk about how I found training data, then did k-folds, then tested the test data on my training model and found an accuracy score of \_\_\_\_\_\_.
  + Freeze last 10 years
  + Shuffle data
* This score was lower than my score just by running an ols and finding a score from there
* Calculate RMSE after un-transforming model
* Talk about how the global population correlates extremely well with predicting the amount of new citizens coming into the US. It shows has a strong correlation, smallest standard error, lowest p-value, and the t-statistic is extremely high thus giving me a strong indication that I can reject the null hypothesis.

How to present this shit:

* Overview
  + Give context
* Data Gathering
  + How all the data came together
* Modeling
  + Explain how OLS fit really well
  + Explain train/test split
  + Elastic Net (say it’s a combo of Ridge and Lasso)
    - Explain all three
    - Think of it as a peanut butter/nutella sandwich
      * Ridge is the peanut butter
      * Lasso is the Nutella
      * Elastic Net is the combo of both
        + You can do each separate, hence them being a PBN and not PBJ because honestly who ever eats just a jelly sandwich, that’s gross
* Analysis
  + Through running the initial OLS model, I found that my variables are extremely correlated from this correlation matrix which shows each variable is correlated with each other. The red being highly correlated, and the light red and light blue not as much.
  + On the right is a graph of the US, this shows how new citizens increased over time.
* Forecast
  + This shows the coefficients of my model along with the accuracy score I received after using an Elastic Net on my train/test split. The training data was 80% of my data, while the test was 20% of my data.
* Potential Improvements:
  + When thinking of what variables I should use, I thought of using new citizens that were refugees. However, that data only dated back for the past 10 years, so I did not use that because it would cut my model down by 80%.
  + The H-1B data would be interesting to me as I would like to see if the US does in fact base it randomly or merit based
* Data
  + These websites are the sites I pulled and scraped data from, ranging from the government to a economic research site
* Tie back to the beginning
  + …Any questions?