Nickel and Dimed: Fluctuating forex rates and how they affect students at Penn

> *Nickel-and-dimed: To be charged many smaller costs or fees that add up to a substantial amount.*

Many international students in the US transfer money from a bank account in their home countries and convert it to USD for their expenditure in the US. However, as the value of foreign currency with respect to USD has been fluctuating and trending downwards in recent times, so has the value of the savings they transfer.

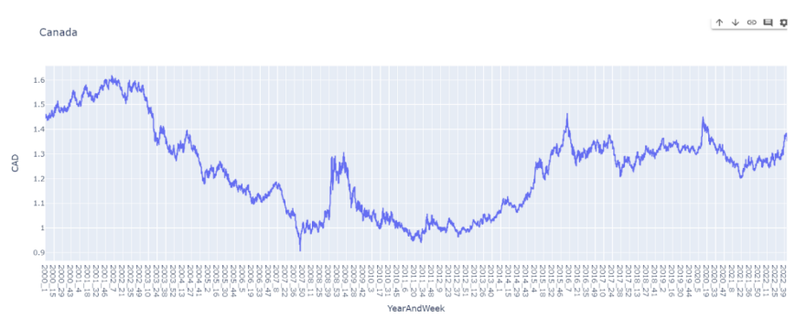
How do we determine whether this phenomenon affects students significantly? Can we quantify it? If it is a real problem, is there any way we can predict Forex trends and help ameliorate the effects?

We explore these questions and more in this article.

. . .

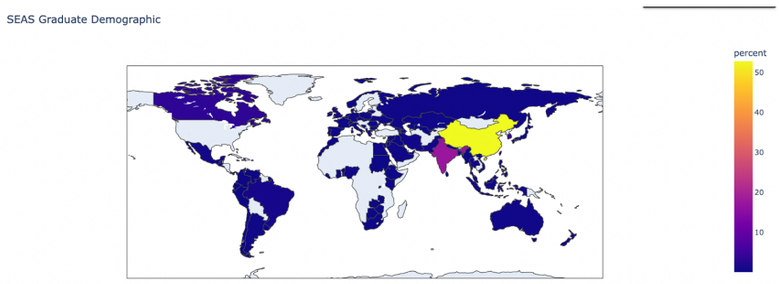
Part 1: Data and design

First and foremost we need Foreign exchange data. We this using the forex-python API (insert link) whose ultimate source is the European Central Bank. We collect daily forex rate data from January, 2000 til October 2022 and aggregate into the mean across weekly tumbling windows, since we believe there is too much noise in daily forex trends to make useful analyses.



Trendline for CAD/USD over the years

International students are the main focus of this project and we acquired aggregate demographic data with respect to SEAS (School of Engineering and Applied Science) students



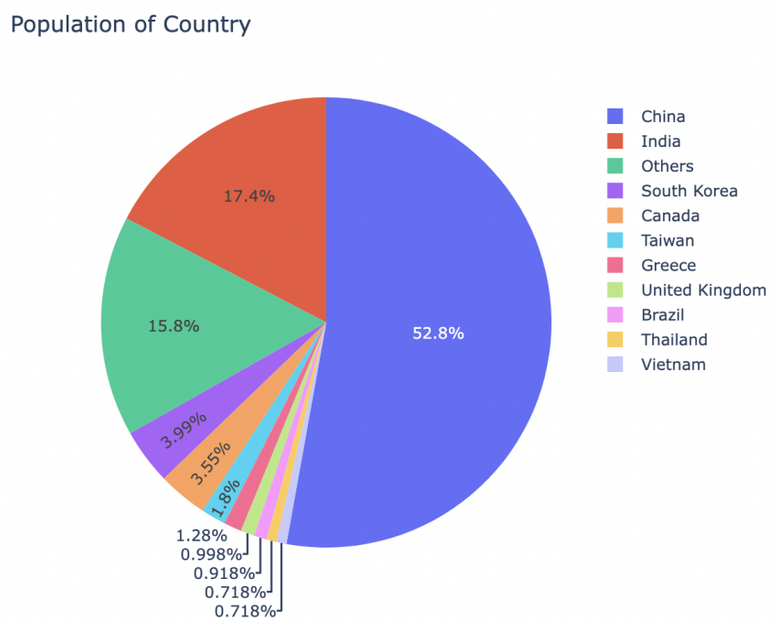
Heatmap for SEAS Graduate demographic

For this project we make use of the top 5 demographics in SEAS. Note: We do not use Taiwan, the 3rd largest demographic, because forex-python does not provide forex information for the New Taiwan Dollar

Finally, as a bonus, we ask the question of whether it is possible to predict forex rates using news articles. Logically, it would seem so , as current events, covered by news, indirectly or directly cause their fluctuations. We attempt to see if it is possible, by scraping financial news article abstracts and headings from the New York Times using the NYT developer API for the same time period as our forex rates and aggregate them on a weekly basis.

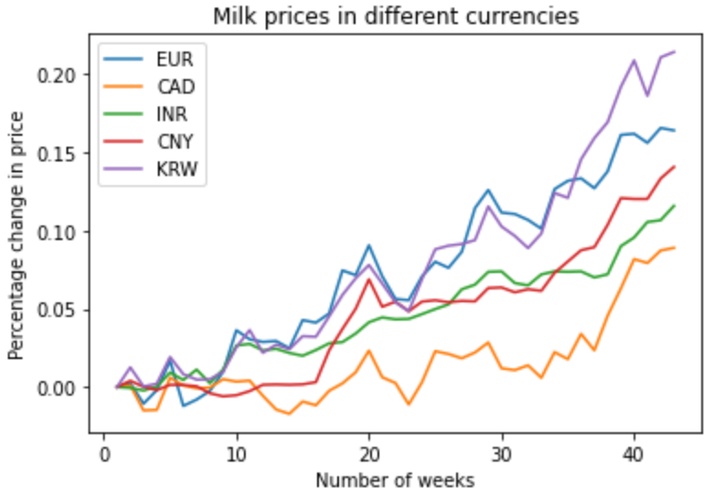
. . .

Part 3: Penn Students and the effect of Forex Trends



% SEAS students belonging to top 10 demographics

The analysis of the data is conducted to figure out how these changes affect the daily lives of students. To illustrate this, we have chosen to track the percentage change in the prices of a commodity, like milk, which is used on a regular basis by students. We have considered the cost of 1 gallon of whole milk for this purpose.



%change in milk prices in top 5 demographics

This graph depicts an upward trend for all countries by approximately 15% over the time period, implying that the cost of milk for an international student has increased over the year.

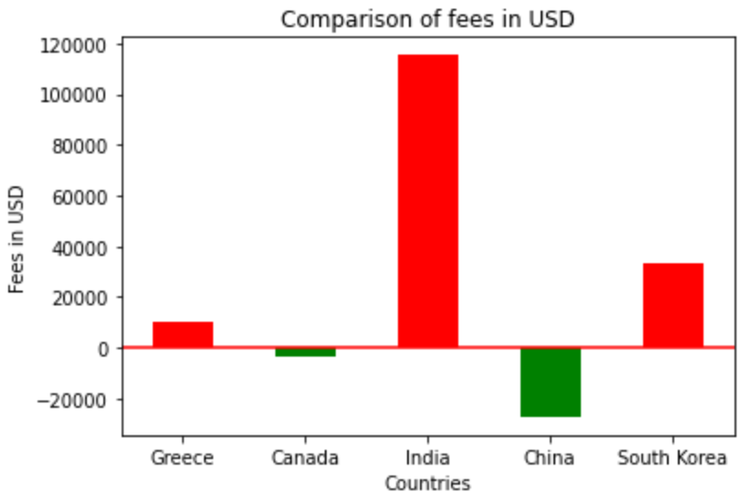
This leads us to think if something inexpensive like milk affects the student's expenditure, then by what margin do overbearing expenses, like fees, have an impact on the student's bank account?

To dig deeper into this question, we compared three types of fee payment time periods. For this, one needs to understand the fee payment methods at UPenn. There are 2 ways to pay your fees: one-time payment and payment plan. One-time payment involves paying the fees at the beginning of the semester where the entire amount due for that semester is paid. The payment plan is paying the fees over the semester at a fixed interval and fixed rate. The entire fee is broken down equally. There are 2 payment plan methods—4 months and 5 months and for the analysis, we have considered the 4 monthly payment plans. To analyze one-time payments, we have considered the first week of January and the first week of September's data. These columns are considered because the semester fees are due during these weeks.

Coming back to the analysis, first, we want to know how much more tuition fee an international student is paying, in their national currency, in the current semester in comparison with the previous semester, if a student is taking part in the one-time payment plan. This gives us a broad overview of the increase in Forex prices. Of the 5 countries taken into consideration, Canada is the only country that did not have to pay anything over what was paid in the previous semester.

To gain insights on a finer time parameter, such as a month, we compare the fees paid over the months of the Spring semester, considering the student has opted for a payment plan. Greece, India, and South Korea have an increase in the amount they have paid each month. For China and Canada, there is only an increment in the 2nd and 3rd months, meaning their currencies did not fall after the third month.

Lastly, we visualize which payment method would help them have more home currency in their hand. The sum of fees paid in a payment plan using their home currency is compared with the fees in the one-time payment for all students in US Dollars.



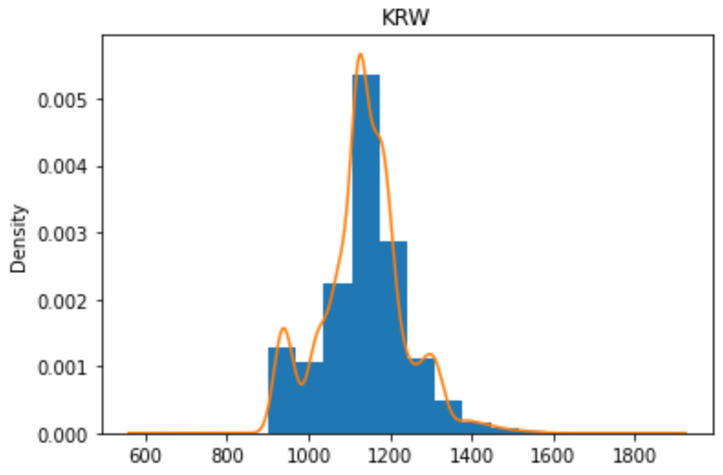
Difference in total fees if all students had paid in one time vs payment plan

For students belonging to India, South Korea, or Greece taking part in the payment plan would have been detrimental since they would have suffered a loss whereas a payment plan would have been preferable for Canadian and Chinese students as they would save up significantly. The bar for India is huge because of the large number of students incoming from India as well as the difference between the currency of India with respect to US.

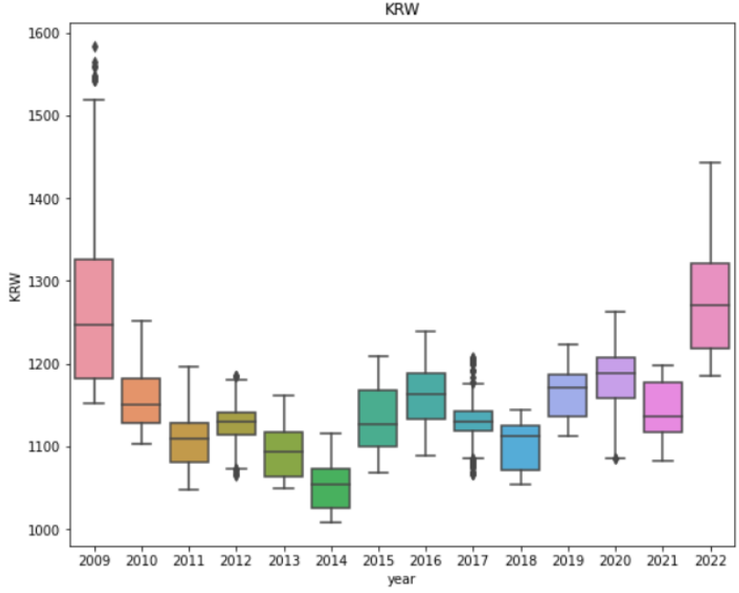
. . .

Part 4: Predicting Forex rates. Can we help our students?

Considering the time series of currencies for each of the top 5 countries UPenn international students belong to, we try to predict the average exchange rates for the currencies in the upcoming week. Since its a time series forecasting problem, we use the relevant visualizations and analysis to get a sense of the data first.

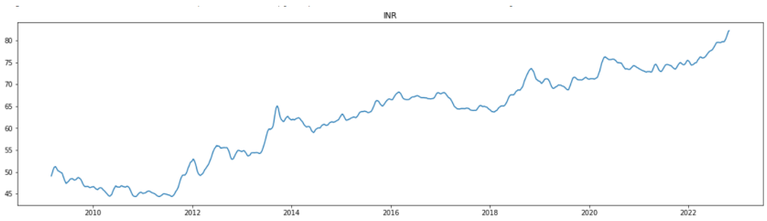


Distribution of South Korean WON/USD



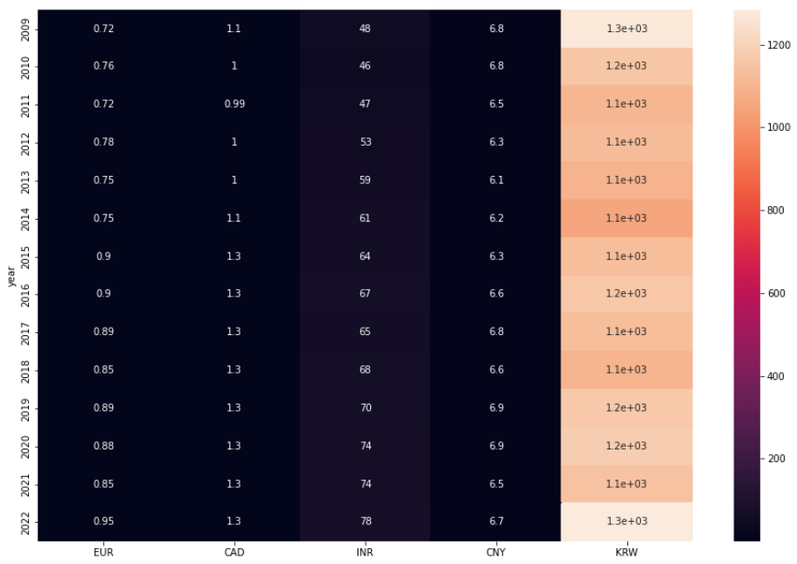
Boxplot of annual Forex fluctuations in South Korean Won over the years

The series showed seasonality and were unfit for modeling directly. For our time series to be stationary, we need to ensure that the rolling statistics remain time-invariant. We further looked at the rolling statistics plots to confirm this. The graph below shows the rolling mean of INR over a sliding window of 30 days. From these visualizations, we inferred that most currencies have non-stationary time series.



INR rolling mean over the years

We tested this hypothesis with the Augmented Dickey-Fuller test, which is a statistical test used to check whether a given time series is stationary or not. Turned out that except for Korea, all other currencies had non-stationary time series. We confirm that the forex rate for these currencies at a timestamp t, is correlated with its values at t-1,t-2...t-n, as shown by the heatmap plotting annual average exchange rate fluctuating year on year for these currencies.



Time varying nature of currencies

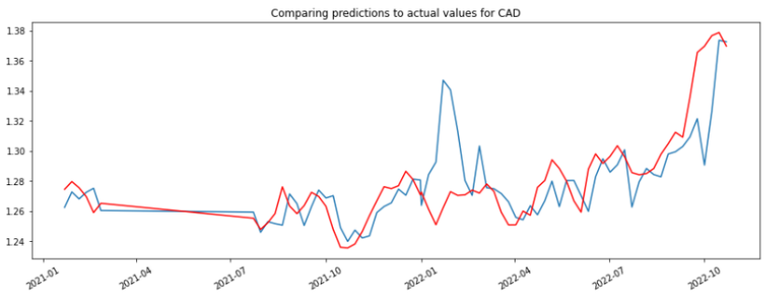
After understanding the data, we decided to use ARIMA for time series forecasting in our case. 'AutoRegressive Integrated Moving Average', is an algorithm that originated from the belief that the past values of a time series can alone be used to predict future values. Although ARIMA handles non stationary time series, we created new smoothened time series by performing transformations for better quality predictions.

To use ARIMA, we need to determine the optimal values of the following parameters:
p: the order of the AR term (Auto-Regressive)
d: the power of differencing
q: the order of the MA term (Moving Average)

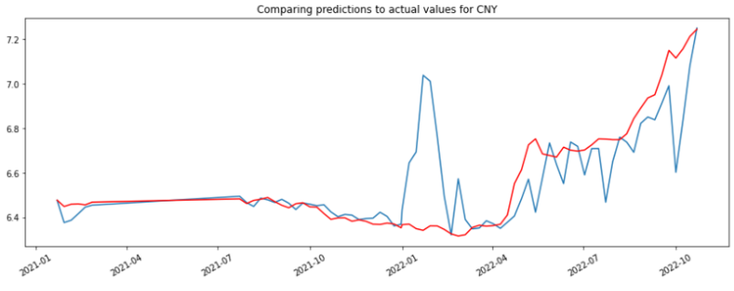
Each time series has its own patterns and it's important to pick a model that identifies those trends and forecasts accordingly. We used iterations similar to grid search to identify the best p,d,q values to train our models for all our series independently.

Next step is to forecast/predict the forex rate. We can either forecast for several days in a chunk, or do a rolling forecast. A rolling forecasting procedure is required given the dependence on observations in prior time steps for differencing and the AR model. Hence, we re-build the ARIMA model after every new observation is received.

Results look aligned when compared with test data:



Predictions on test data (CAD)



Predictions on test data (CNY)

Our models were able to predict forex values for these currencies with a significantly low RMSE and AIC (Akaike Information Criteria). However, we observe from the charts that models can perform better terms of predicting sudden peaks and troughs. For the future, we aim at improving our models further by incorporating different techniques like SARIMAX in order to read and use information like seasonal cycles and exogenous factors from the data, and use them for training our models, leading to finer predictions.

Now that we have predicted the future forex rates for our currencies, let's also see how this can help international students from these countries.

We try to assess the impact of weekly changes in the exchange rates of these countries on the monthly house rent that international students pay (assuming avg as $1000 in Philadelphia). Our model can help students decide if they should withdraw and convert their currency this week or the following week in order to save money.

Based on our predictions after running the model considering 10/22/2022 (last weekly window in our dataset) as the current week and the week following it as the next week, we can say that students from Greece, India and China should pay their rent this week as exchange rates for these currencies are predicted to go up. Students from South Korea and Canada can wait and pay the rent next week as exchange rates for these currencies are predicted to go down.

This way, our model could have helped save $6604 net (profit-loss) for all international students overall in a week!!
Profit—Money saved by students if our model predicts the right direction of forex rates.
Loss—Money lost by students if our model predicts the wrong direction of forex rates.

. . .

Part 5: Predicting Forex rates from text

While the above sections should have convinced you that Forex trends may indeed have an effect on international students at Penn, we also want to see if it is possible to predict forex rates based on news articles.

Let's dive in!

First off we need to define our tasks clearly.

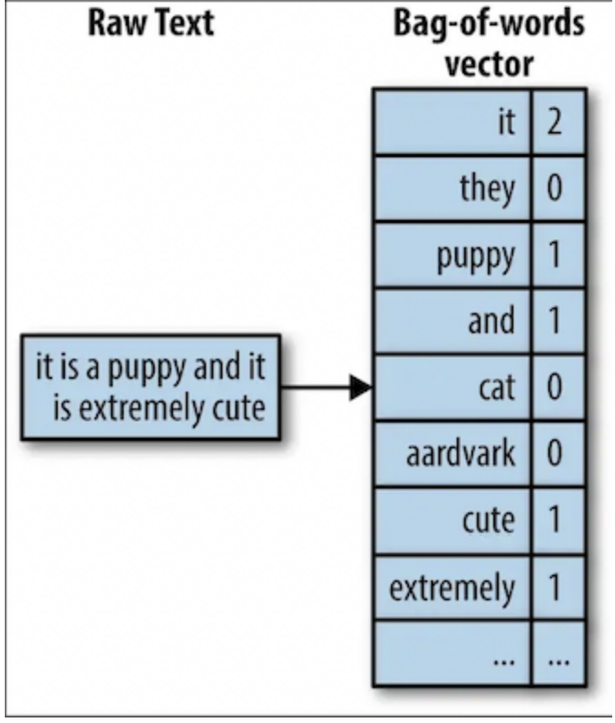
i) Classification:
Given the news articles for a given week, as in the previous section, we want to predict whether the average forex rate across the next week will increase or decrease. I.e go 'Up' or 'Down'. This is a binary classification problem.

ii) Regression:
Given the news article for a particular week we want to predict the forex rate value for the next week.

Feature Extraction

For each week, we have a list of text articles. After preprocessing, which included removal of punctuation, converting to lowercase and stemming, we try to extract features in two ways. We run all our experiments with each of these methods:

i) Bag Of Words



Bag of Words encoding sample

Each feature in a Bag of Words representation is the number of times each token has appeared in the text (term frequency). It has a sparse representation as many features are zero

ii)Tf- Idf:

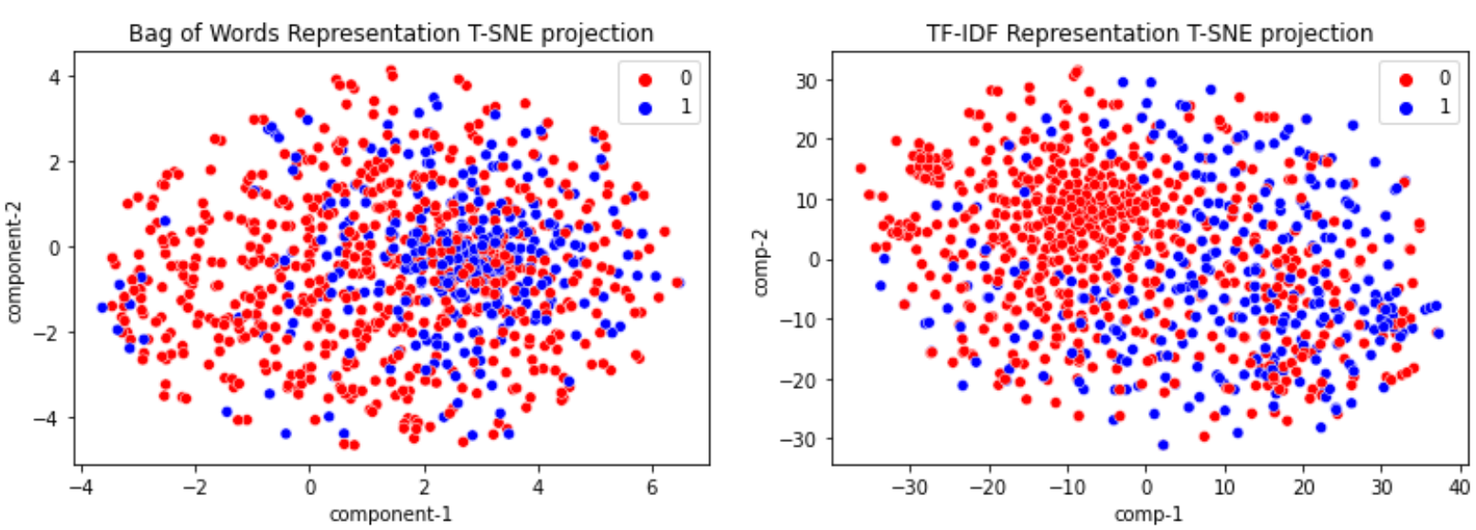
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Tfidf formula

Tf-Idf is similar to bag of words but each term is also weighed according to its importance.

Experiments and Results:
For simplicity, we perform all these experiments on the exchange rate of INR.
Here is what happens if we attempt to plot our extracted features in 2D vector space.



t-sne plots for feature vectors in 2D space

Text is very complex and even though there isn't any clear separation between the two classes in 2D vector space it is possible to see that our blue class (Up) and red class(Down) points are concentration in different regions, which itself is very interesting. This gives hope that further separation is possible in nD space/

1. Classification:

We tried out 4 models:

a) Logistic Regression with No Penalty

b) Logistic Regression with L1 penalization

c) Logistic Regression with L2 penalization

d) Random Forest

Results

Search this file…

	Model Type	Accuracy
2	Logistic Regression without Penalty	73.64%
3	Logistic Regression - L1 Penalty	74.06%
4	Logistic Regression - L2 Penalty	71.97%
5	Random Forest	69.03%

BoW Classification Results.csv hosted with ❤ by GitHub view raw

BoW Classification results

Search this file…

	Model Type	Accuracy
2	Logistic Regression without Penalty	72.8%
3	Logistic Regression - L1 Penalty	71.13%
4	Logistic Regression - L2 Penalty	71.15%
5	Random Forest	69.46

Tf-Idf Classification Results.csv hosted with ❤ by GitHub view raw

Tf-Idf classification results

Here are some interesting things we note:

1. **For BoW representation, Logistic Regression with L1 penalty performs best.** This makes sense as BoW form is a sparse form of representation, certain individual terms probably have more feature importance, and L1 penalization pushes non-important features to zero.

2. **For Tf-Idf the non-regularized model performs best.** This probably requires more inspection as this generally does not happen

3. Random forest performs poorly for both representations

And perhaps most importantly:

4. **Our classifier performs quite well!** Simple models with minimal parameter tuning are able to predict whether the forex rate for the next week

2. Regression:

We consider 3 model variants:

1. Linear Regression
2. LASSO regression
3. Ridge Regression

Results:

```
Model Type, Pearson Correlation Coefficient
Linear Regression, -27.09
LASSO Regression, -13.90
Ridge Regression, -18.42
```

BoW regression results hosted with ❤ by GitHub view raw

BoW Regression Results

```
Model Type, Pearson Correlation Coefficient
Linear Regression, 0.68
LASSO Regression, 0.59
Ridge Regression, -0.006
```

Tf-Idf Regression results hosted with ❤ by GitHub view raw

Tf-Idf Regression Results

Interesting observations:

1. Regression using Tf-Idf features outperforms regression with BoW features by far
2. Once again, for Tf-Idf, regression without penalty outperforms penalized linear regression
3. It's very interesting to note that simple linear regression is almost linear correlation (R^2 ~= 0.7)between text features and Forex rates.

We see that even simple models with minimal fine-tuning seem to be able to find a reasonable correlation between news article data and forex rates for the next week. This leads us to believe that it may indeed be possible to predict Forex with good accuracy with just News data.

. . .

Part 6: Challenges

1. Data collection and integration—Collecting and combining different types of data from multiple sources presented multiple

preprocessing obstacles. This led us to put in a significant chunk of time in data cleaning and exploratory data analysis.

. . .

Part 7: Future Work

1. Combining NLP and time series models to enhance our predictions—Ensemble techniques can be used to build a model on top of the two different approaches we explore in this project. This will be a great way to increase accuracy of our predictions.

2. Exploring more cases where forex fluctuations impact international students

3. Transitioning to an actionable —In this project we tried to highlight a noteworthy problem, in future we would like to present a solution to it using more interesting Data Science techniques.

4. The discussed results for NLP models are for minimal finetuning, we can work on improving this.

. . .

Part 8: Conclusion

In this work we attempt to analyze whether fluctuations in Forex rates affect international students enrolled under the SEAS department at Penn. We do this by looking at the top student demographics and considering specific scenarios with respect to Forex rate and their expenditure, such as paying fees all at once vs in parts in the form of a payment plan, expenditure for daily items such as milk. We attempt to model these forex rates as a time-series using ARIMA, train a model, and show how using our model could save students money. Finally, we explore whether we can predict forex rates using News headlines and train surprisingly well performing models.

In light of this work it is reasonable to conclude that fluctuating forex rates affecting international students is a noteworthy problem.