Define problem statement

Identify Data Sources

Clean Data

Discuss results and analysis

Implement ML algorithm(s)

Identify one or more ML algorithms

Feature selection

Conclusion

Data Cleaning steps:

S&P data:

* Take S&P 500 Historical data\_Start till 2018\_csv from datahub.io \cite{website12}

This gives S&P500 data till 2018. Has 1768 rows

Attributes: Data, SP500, CPI, Long Interest Rate

* Take S&P 500 Historical Data.xls file from Investing.com \cite{website11}

Sort the file by date from old to new. Took 59 records from this data set

Take S&P500 data from 5/1/2018 till 2023 from this file

Attributes: Price

* Random sampling to check the consistency of between data sets. The Price in the First data set lies within the High and Low prices in the second data set for each month. Hence good to use

GDP data:

* Take GDP data set from Fed \cite{website13}. This gives monthly GDP from 1992

Attribute: Nominal GDP

* Change all date formats on all Data sets to be consistent MM/DD/YYYY
* Take Pre 1992 GDP data from the website \cite{website14}

Change the Date format on this data set to match the date format on the Consolidated Data set. YYYY-MM-DD is changed to MM/DD/YYYY

Do Vlook up on date from the Consolidated Data set and obtain the GDP numbers from this data set

* Random sampling to check the consistency of between data sets. The Price in the First data set is pretty close to that from the second data set for each month. Hence good to use

Fed rate:

* Take the Fedfunds rate data from \cite{website06}

Do Vlook up on date from the Consolidated Data set and obtain the Fed rate numbers from this data set

Inflation data:

* Use the Historical Inflation data obtained from BLS
* This data needs significant transformation

Unemployment data:

* Use the Historical Inflation data obtained from BLS
* This data needs significant transformation

MACD:

* Plug in the Data and SP500 values from the Consolidated Data set into the MACD template
* Do Vlook up on date from the Consolidated Data set and obtain the MACD numbers from this data set

RSI:

* Plug in the Data and SP500 values from the Consolidated Data set into the RSI template
* Do Vlook up on date from the Consolidated Data set and obtain the RSI numbers from this data set

=IF(A2="Jan",1,IF(A2="Feb",2,IF(A2="Mar",3,IF(A2="Apr",4,IF(A2="May"=5,IF(A2="Jun",6,IF(A2="Jul",7,IF(A2="Aug",8,IF(A2="Sep",9,IF(A2="Oct",10,IF(A2="Nov",11,IF(A2="Dec",12,""))))))))))))

**Data Cleaning:**

Missing Data: Deleted rows from 1871 to 1957 because no data was available for GDP, Fed rate, Inflation and Unemployment. Not even enough data to fill in missing values. Left with 783 rows

Fed funds rate for 3/1/23 was missing. However, Fed announced a rate hike of 0.25% on 3/22/23. So added .25 to the previous value from 2/1/23

Inflation rate for 3/1/23 was missing. Inflation has remained between 5.5-6% for the past couple of months. Hence, an average of the past four month inflation rates would be a close approximation for the Inflation rate for March 2023.

Unemployment data hasn’t been published rate for 3/1/23.

Unemployment rate has remained between 3.5 – 3.7% for the past couple of months. Hence, an average of the past four month inflation rates would be a close approximation for the Inflation rate for March 2023.

GDP numbers for the initial years are available for each quarter.

Month 1 GDP = GDP1

Month 2 GDP = Not available

Month 3 GDP = Not avai

Month 4 GDP = GDP4

Month 2 GDP calculated as GDP1 + (GDP4-GDP1)/3

Month 3 GDP calculated as GDP1 + 2\*(GDP4-GDP1)/3

This gives a good approximation of the missing GDP values

Transformation: Date formats for each data set had to be transformed

Merge/Consolidate similar columns:

* Multiple columns with SP500 data is consolidated into one column
* Multiple columns of GDP data is consolidated into one column
* Multiple columns of Fed funds data is consolidated into one column

Merge the 12 CSV files into one consolidated CSV

1827 records

Standardize data in each data set. Sort records by date and format date in mm/dd/yyyy format

12 CSV files

Cleaned Data set with 783 rows

Fill in missing values in GDP column using formula

Remove rows from 1/1/1871 to 1/1/1958

**Exploratory Data Analysis:**

**Part 1**

1. What is exploratory data analysis? Why is it essential in a data science or data analytics project?

Exploratory Data Analysis (EDA) is data exploration technique to understand various aspects of data. IT is a process of investigating dataset to uncover patterns, anomalies etc and understand the dataset better. The relationship between the data variables is understood through EDA. Insights are gathered about the data during this phase before moving on to more complex processes such as using algorithms/ predictive modelling etc in the Data analysis/ Data science life cycle. This is a method used to analyze and summarize data sets

EDA helps in identifying faulty points in Data even after the Data cleaning phase. Data can be further cleaned during this process with a true understanding of the relationship between variables. Which gives a true understanding of data

1. Understand the data (variables, number of columns) etc
2. Clean the data -> Already done in the previous phase
3. Analyze relationship between variables
4. What are the various exploratory data analysis techniques? What specific techniques work for your project?
   1. Clustering and dimension reduction techniques
   2. Univariate visualization of each field in the raw dataset
   3. Bivariate visualizations and summary statistics to assess the relation between the dependent and independent variables
   4. Multivariate visualizations to understand interactions between different fields in the data
   5. K-means clustering: Data points are assigned into K groups. The data points closest to a particular centroid will be clustered under the same category. This method is used in market segmentation, pattern recognition, image compression
   6. Predictive models, such as linear regression and statistics

Techniques mentioned in #2, #3, #4 seem to be applicable for my project

1. Explain the details of the techniques and the results of your exploratory analysis. (You may include a diagram to explain your exploratory analysis phase.)
   1. Imported the necessary libraries (numpy, pandas, matplotlib, seaborn) and read the data set into a data frame
   2. Displayed the head (first 5 rows)
   3. Used df.shape to display the total rows and columns of the data frame (783 rows and 8 columns)
   4. Displayed the data types of the columns using df.info()
   5. Converted the Date format from object to date using

df["Date"] = pd.to\_datetime(df["Date"])

Now all the columns are in float format

* 1. Described the data (Counts, Median, Max, Min etc) using df.describe()

Table

Description automatically generated

* 1. Checked for Null values using df.isnull().sum()

No data is null. So concluded that the data is clean

Text

Description automatically generated

* 1. Checked for Unique values using df.nunique()

Most rows are unique

Text, letter

Description automatically generated

* 1. **Univariate visualizations:** 
     1. Generated Histograms for each variable to understand distribution for data
     2. Generated Box plots for each variable
  2. **Bivariate visualizations:** Generated Scatter plots for each independent variable on X axis and the dependent variable (SP500) on the Y axis
  3. **Multivariate Visualizations**:
     1. Relationship analysis: Did correlation analysis and generated Heat maps

corelation = df.corr()

sns.heatmap(corelation, xticklabels=corelation.columns, yticklabels=corelation.columns, annot=True)

Chart

Description automatically generated

This shows that that there is not much correlation between the independent variables. So all the columns can be used and no need to delete any

* + 1. Generated Pair plots

sns.pairplot(df)

Diagram, engineering drawing

Description automatically generated

1. What insights are you learning from this particular phase?

I began getting a vision of the process/path towards the end goal. The data which was all numbers during the earlier stages is beginning to make sense. I got a picture of the boundaries of each variable, what values for each value seem normal (more frequently occurring), what values are extreme on both sides etc. I got a good understanding of the relationship between the variables. Some relationships are intuitive but some relationships are not, like the relationship between GDP and Inflation, Inflation and unemployment, Feds rate and unemployment etc. Overall, the exploratory data analysis phase gave me an idea of how to proceed next

**Part 2**

1. Clickable link to your GitHub repo: <https://github.com/pamidisushma02/Sushma_SP500Prediction_Capstone_Project>
2. Clickable link to your Overleaf report: <https://www.overleaf.com/read/jpdsbxmjtyvq>

**Prediction:**

RNN, LSTM

<https://www.youtube.com/watch?v=lncoLfue_Y4>