Data Analytics Final Project

Contents

- Selected topic
- Reason the topic was selected
- Description of the source of data
- Questions the team hopes to answer with the data
- Description of the data exploration phase of the project
- Description of the analysis phase of the project
- Technologies, languages, tools, and algorithms used throughout the project
- Result of analysis
- Recommendation for future analysis
- Anything the team would have done differently

Project Outline

- **Topic**
  - Creating a predictive algorithm tailored to recession duration
- **Why Selected?**
  - Relates to current events and could have an impact on determining the duration of a recession due to Covid-19
- **Data sets provided by Federal Reserve Bank of St. Louis and WSJ**
  - See next slide for more details.
- **Question that we hope to answer**
  - Can you predict the length of this current downturn due to the Pandemic by looking for clues in economic and market data from our historical periods?

Source Data Sets

1. Dates of U.S. and international recessions as inferred by GDP-based recession indicator
    a. https://fred.stlouisfed.org/series/JHDUSRGDPBR
    b. Hamilton, James, Dates of U.S. recessions as inferred by GDP-based recession indicator [JHDUSRGDPBR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/JHDUSRGDPBR, May 27, 2020.
2. Index and ETF Price Information - WSJ
    a. Data set provides open, high, low, and close price information.

Preliminary Features

- Feature Selection is a critical component in a Data Scientist’s workflow. When presented data with high dimensionality,
  - Training time increases exponentially with number of features.
  - Models have increasing risk of overfitting with increasing number of features.
- Options for handling missing data
  - Do nothing
  - Drop the row that has the missing value
  - Fill in the row that has the missing value.
- We used the `dropna()` method to drop missing data, and the `drop()` method to drop country column since we are only analyzing the US at this point.

Machine Learning Dataset Info

- Training dataset was 75% of the dataset
- Testing dataset was 25% of the dataset

Machine Learning Model

Choice and Benefits

Random forest algorithm will sample the data and build several smaller, and simpler decision trees. Each tree is simpler because it is built from a random subset of features. Random forest algorithms are beneficial because they:

1. Are robust against overfitting as all of those weak learners are trained on different pieces of the data.

2. Can be used to **rank the importance** of input variables in a natural way.

3. Can handle thousands of input variables without variable deletion.

4. Are robust to outliers and nonlinear data.

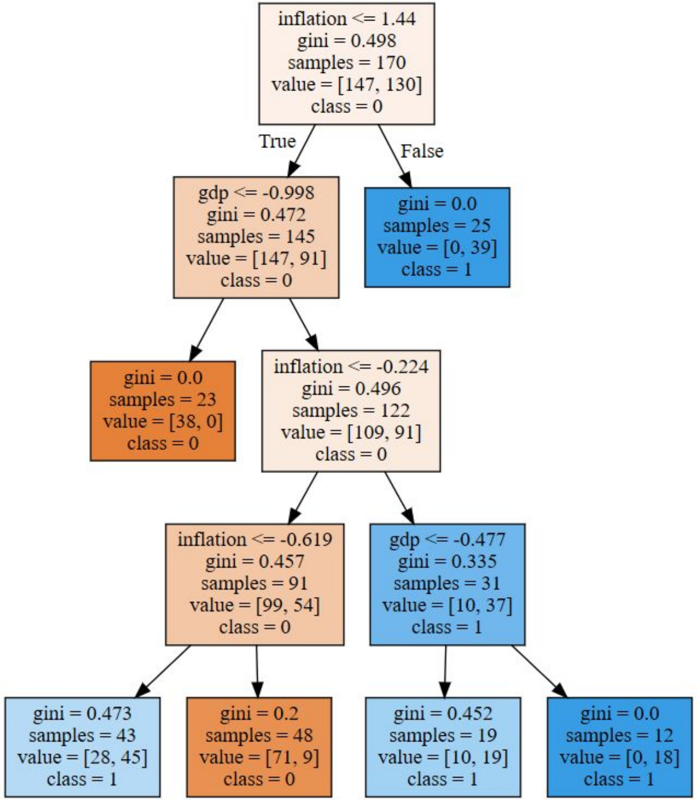
5. Run efficiently on large datasets.

Machine Learning Model Cotd.

Limitations

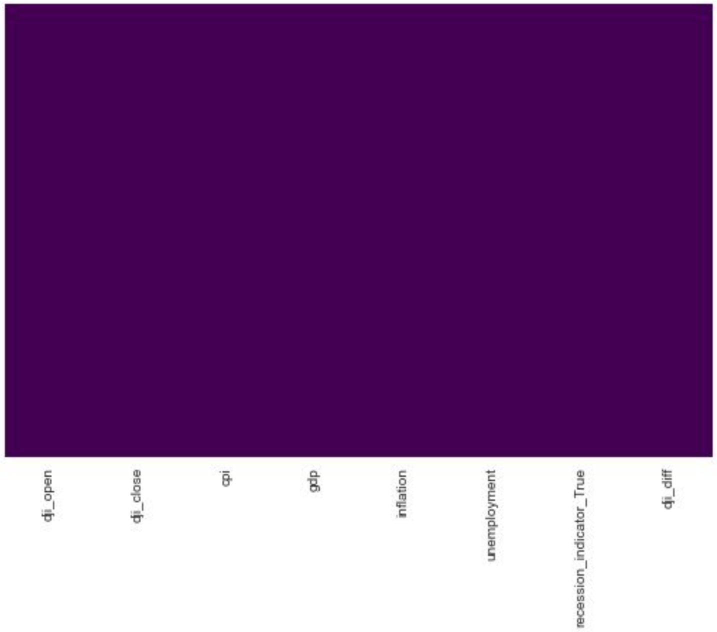
The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. Random Forest creates a lot of trees and require much more time to train.

Random Forest Visualization



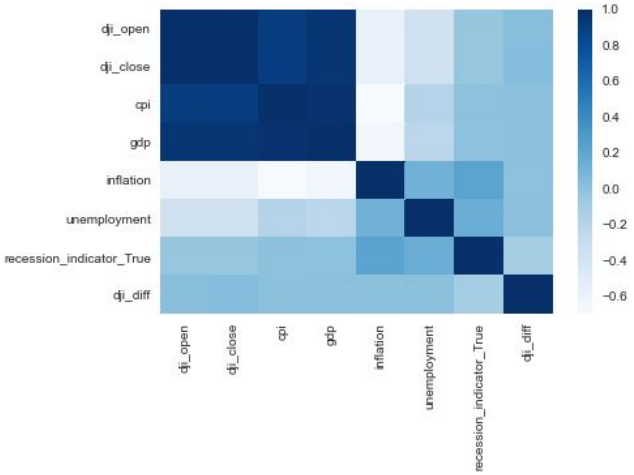
Exploratory Data Analysis

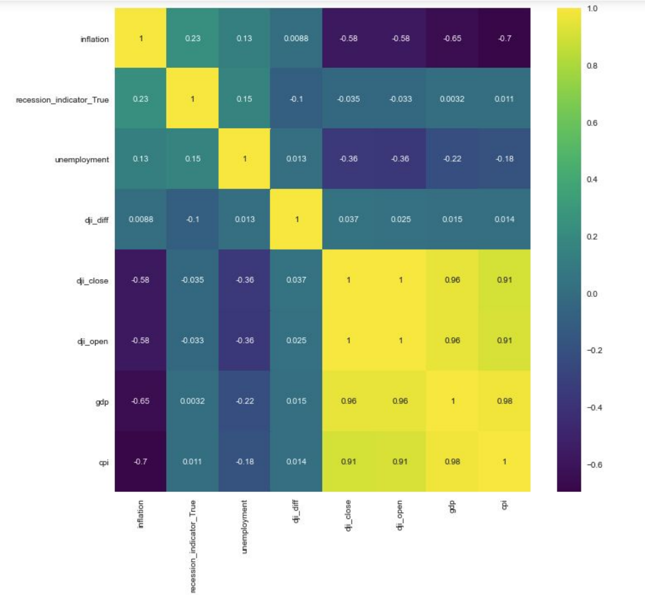
- Checked for missing values in data using heatmaps



Exploratory Data Analysis (cont)

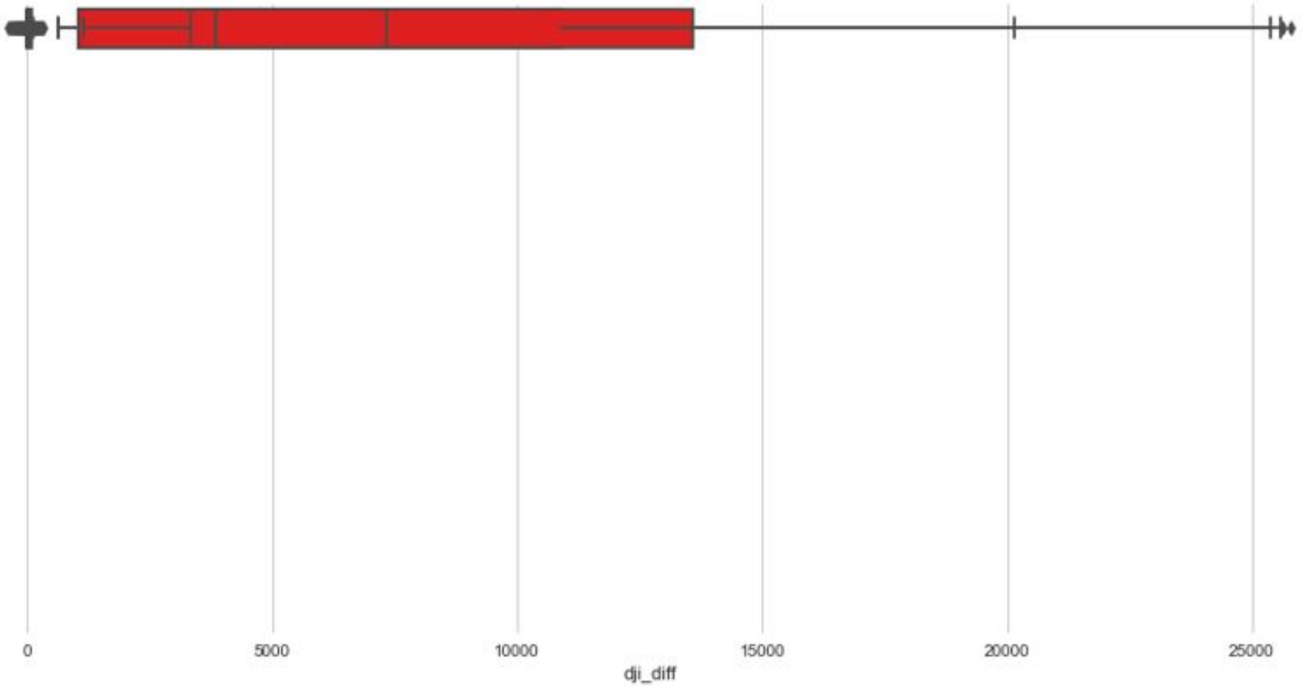
- Assessed Correlation between variables using correlation matrix





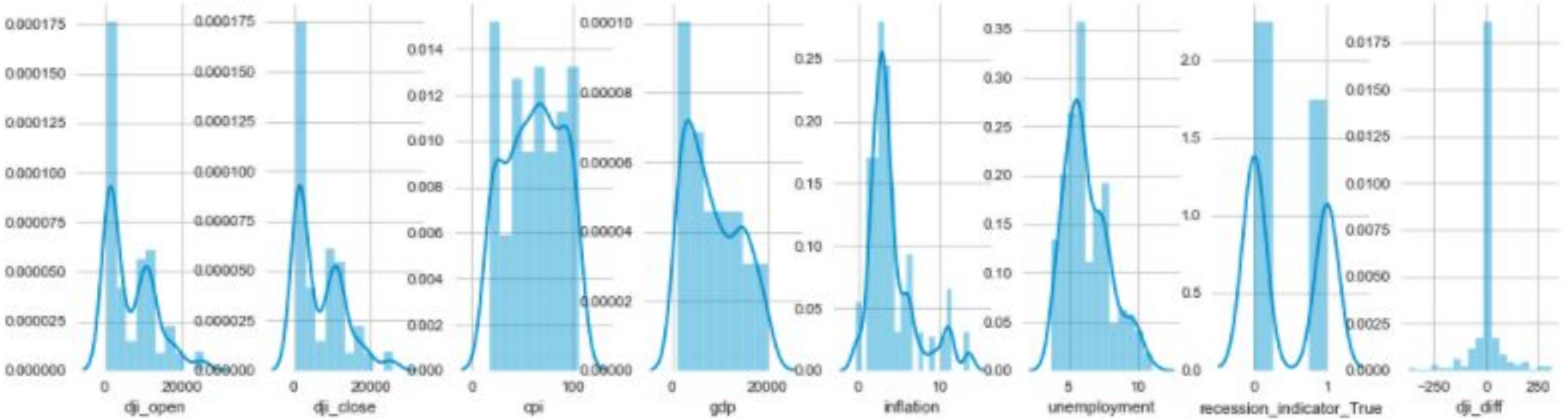
Exploratory Data Analysis (cont)

- [



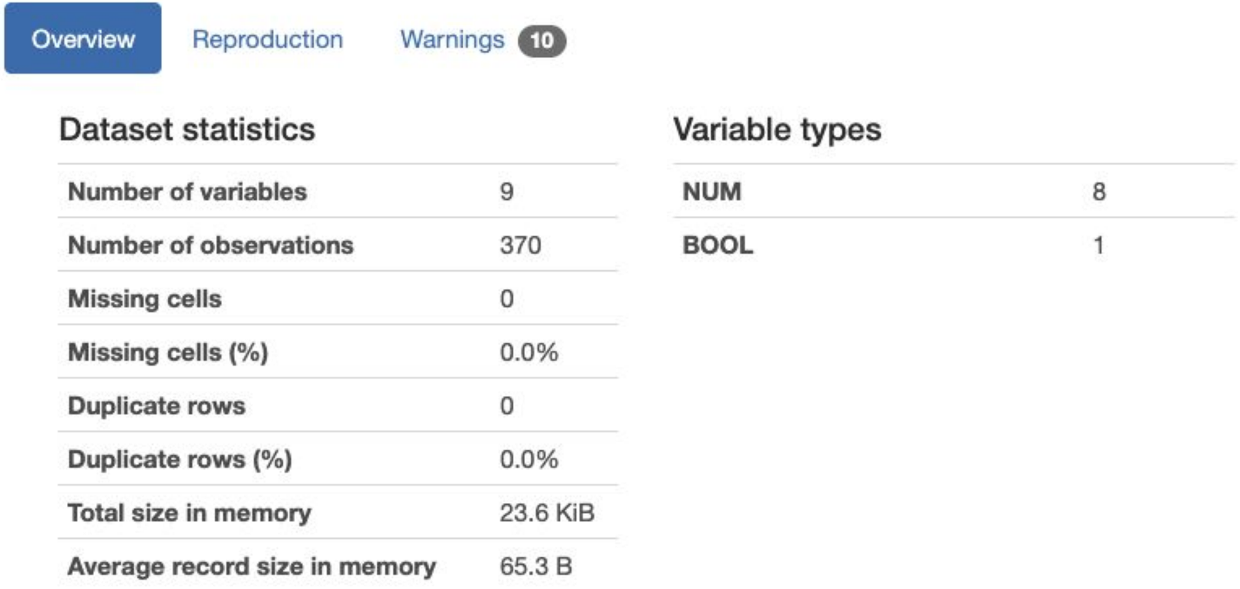
Exploratory Data Analysis (cont)

- Checked for distribution skewness



Exploratory Data Analysis (cont)

- Used a tool called “Pandas Profiling”



Dataset statistics

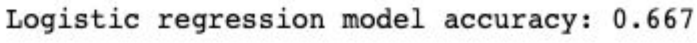
Number of variables	9
Number of observations	370
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	23.6 KiB
Average record size in memory	65.3 B

Variable types

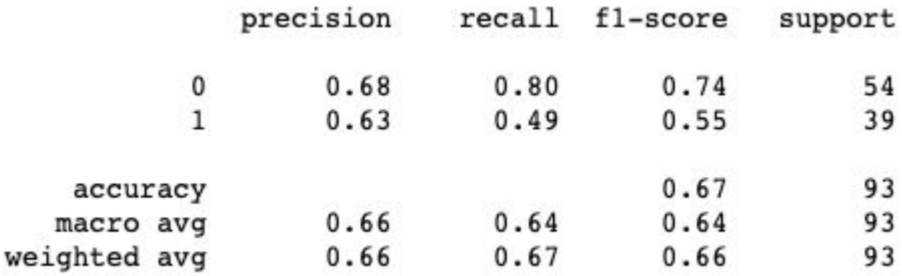
NUM	8
BOOL	1

Analysis Phase

Logistic Regression Model Results



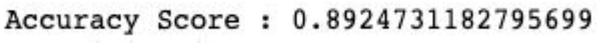
```
Logistic regression model accuracy: 0.667
```



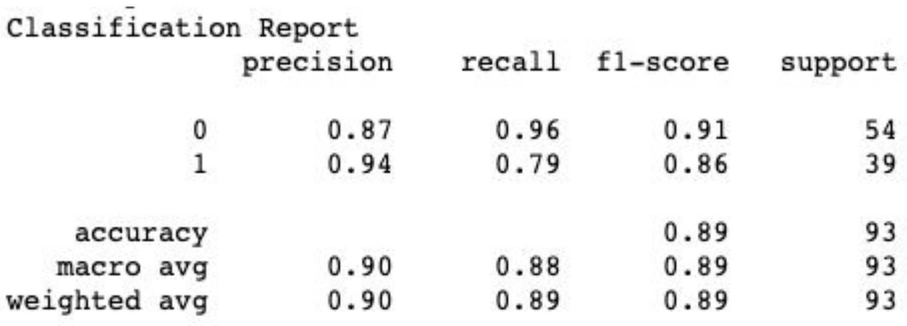
```
              precision    recall  f1-score   support

           0       0.68      0.80      0.74        54
           1       0.63      0.49      0.55        39

    accuracy                           0.67        93
   macro avg       0.66      0.64      0.64        93
weighted avg       0.66      0.67      0.66        93
```

Analysis Phase

Random Forest Results



```
Accuracy Score : 0.8924731182795699
```



```
Classification Report
              precision    recall  f1-score   support

           0       0.87      0.96      0.91        54
           1       0.94      0.79      0.86        39

    accuracy                           0.89        93
   macro avg       0.90      0.88      0.89        93
weighted avg       0.90      0.89      0.89        93
```

Next Steps

As a result of our analysis, we learned:

1. Initially had issues with the dataset, where the dataset was too small and our Random Forest accuracy was 100%, a clear indication of overfitting.
2. Adjusted the dataset to make it larger and the Random Forest accuracy dropped to 89%.
3. In looking at the correlation data, we are still concerned that we may be overfitting the data. In module 3 we want to look at Time Series data to see the results.

Technologies

Data Cleaning and Analysis

- Python Pandas library will be used to clean, prepare and explore the data and perform the initial analysis; potentially to fill in/ drop any NaN data, remove redundant columns, create binning, etc.

Database Storage

- Postgres is the database we intend to use for storing the data. Click this link, https://github.com/UCB-Extension-Team-6-Final-2020/Predictive-Market-Analyzer/blob/master/images/DB-ERD.png, for the database design.

Dashboard

- We will utilize Tableau for our dashboard to create visuals for data storytelling.
- It will be hosted on Tableau Public.

Technologies Cotd.

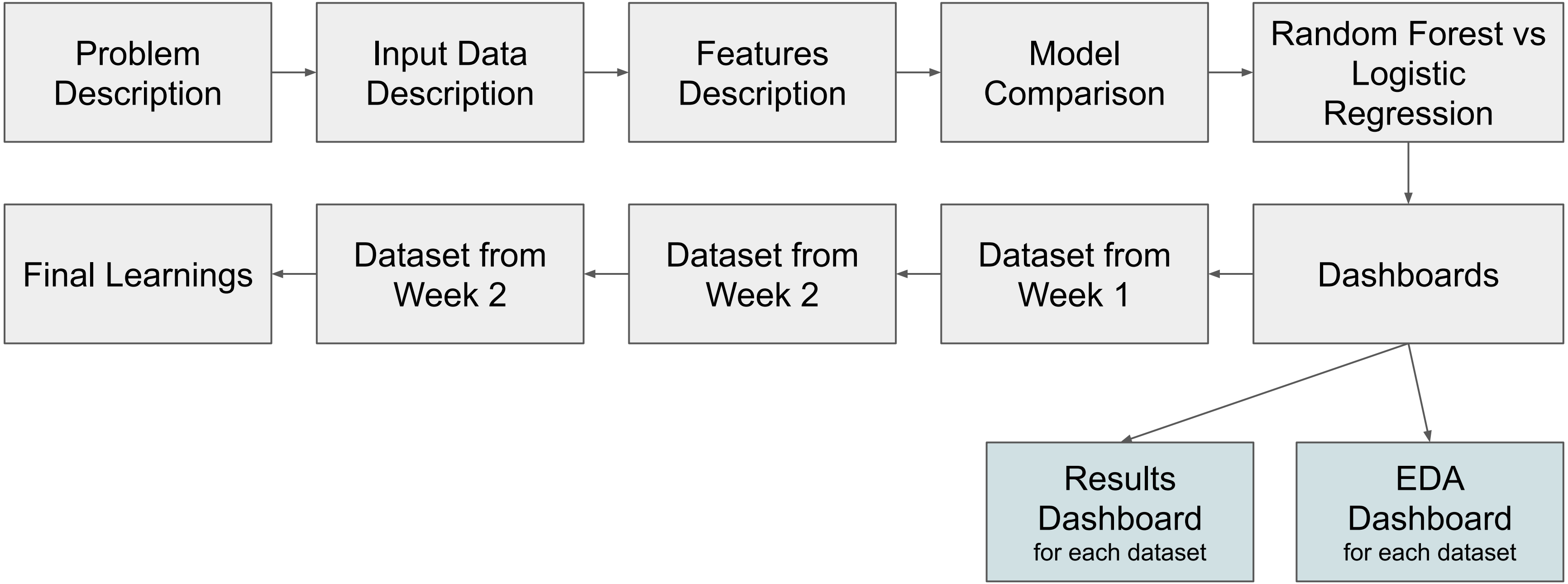
Machine Learning

- SciKitLearn and Tensorflow are the Machine Learning libraries we'll be using.

The process -

- Use One-Hot Encoder to understand and evaluate any categorical variables,
- split our preprocessed data into features and target arrays, and then the training and testing dataset,
- scale the data,
- perform either Logic Regression, Support Vector Machine (SVM), or Random Forest,
- compare with Basic Neural Network (1 hidden layer), and Deep Learning Model Design (2 hidden layers)

Storyboarding



Problem Description

Input Data Description

Features Description

Model Comparison

Random Forest vs Logistic Regression

Dashboards

Dataset from Week 1

Dataset from Week 2

Dataset from Week 2

Final Learnings

Results Dashboard for each dataset

EDA Dashboard for each dataset

Results Dashboard

Accuracy

___ %
Logistic Regression

___ %
Random Forest

Precision

___ %
Logistic Regression

___ %
Random Forest

Sensitivity

___ %
Logistic Regression

___ %
Random Forest

Feature Correlation

Correlation Heatmaps

Recession Prediction

Line graph showing historical vs forecasted

EDA Dashboard

DF Shape

____
Observations

____
Features

Null Values

Bar Chart by
Feature (Column)

Variable Types

____
Number

____
Boolean

Skewness

Bar/Line Chart by Feature (Column)

Outliers

Boxplot

Line Graph of Each Feature