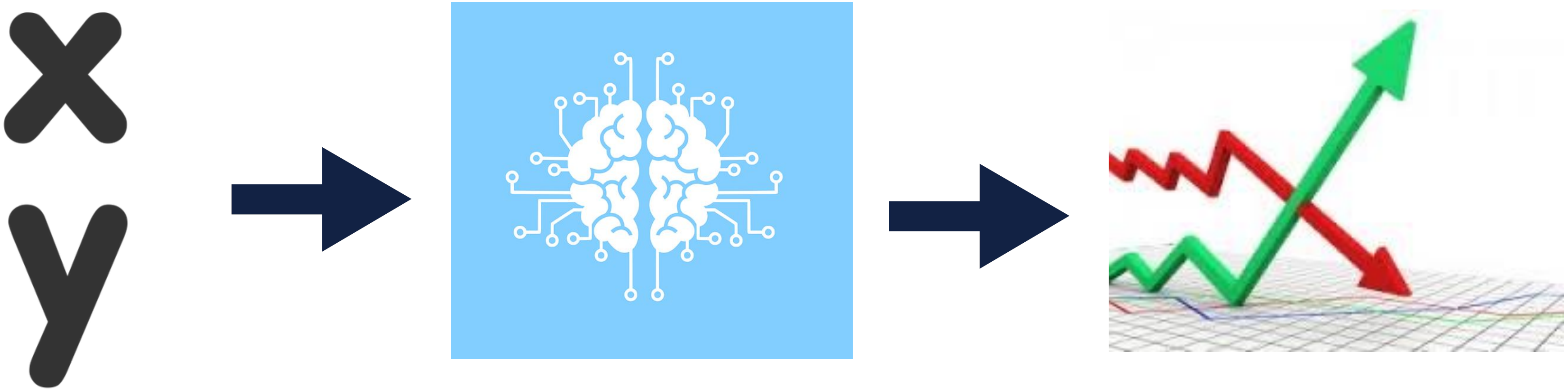
Predicting Closing Stock Exchange Prices on the NYSE & NASDAQ

Matthew Griswold, Andrew Mekhail, Michal Michael, Shelley Mitchell, Dhaval Patel

Big Question

Based on a given set of variables, can machine learning predict whether a stock exchange index will close higher or lower than previous close?



Applicability

Personal Finance

acorns
Invest the Change

charles
SCHWAB

SoFi

Mobile Investing

Robinhood

TD
Ameritrade

Project Background

Dataset

We utilized two datasets from Kaggle (please see link below)

https://www.kaggle.com/mattiuzc/stock-exchange-data

Columns

The two CSVs were cross referenced by the Index to allow for more detail.

- **indexProcessed.csv** - Stock Exchange Index, Date, Open, High, Low, Close, Adj Close, Volume
- **indexInfo.csv** - Region, Exchange, Index, Currency

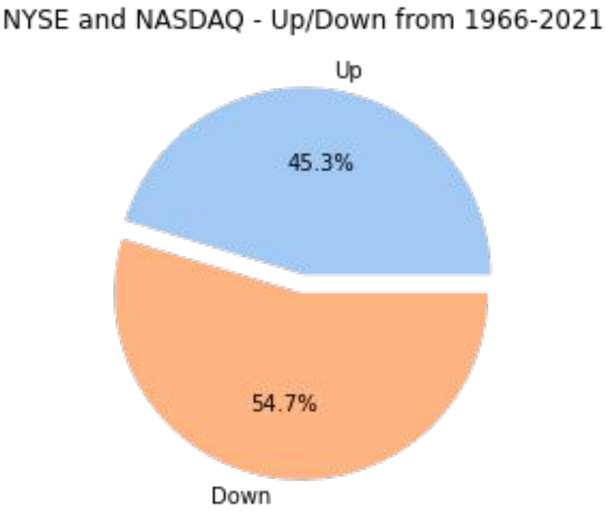
Technology

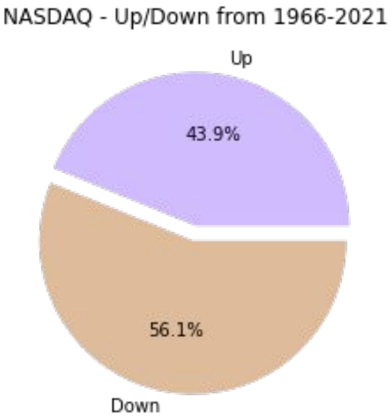
We used 3 main technologies for different aspects of the project:

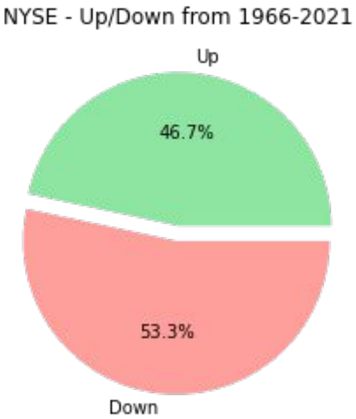
- **SQL - Postgres** - database
- **Python** - data preparation and machine learning
- **Tableau** - data visualization

Exploratory Analysis

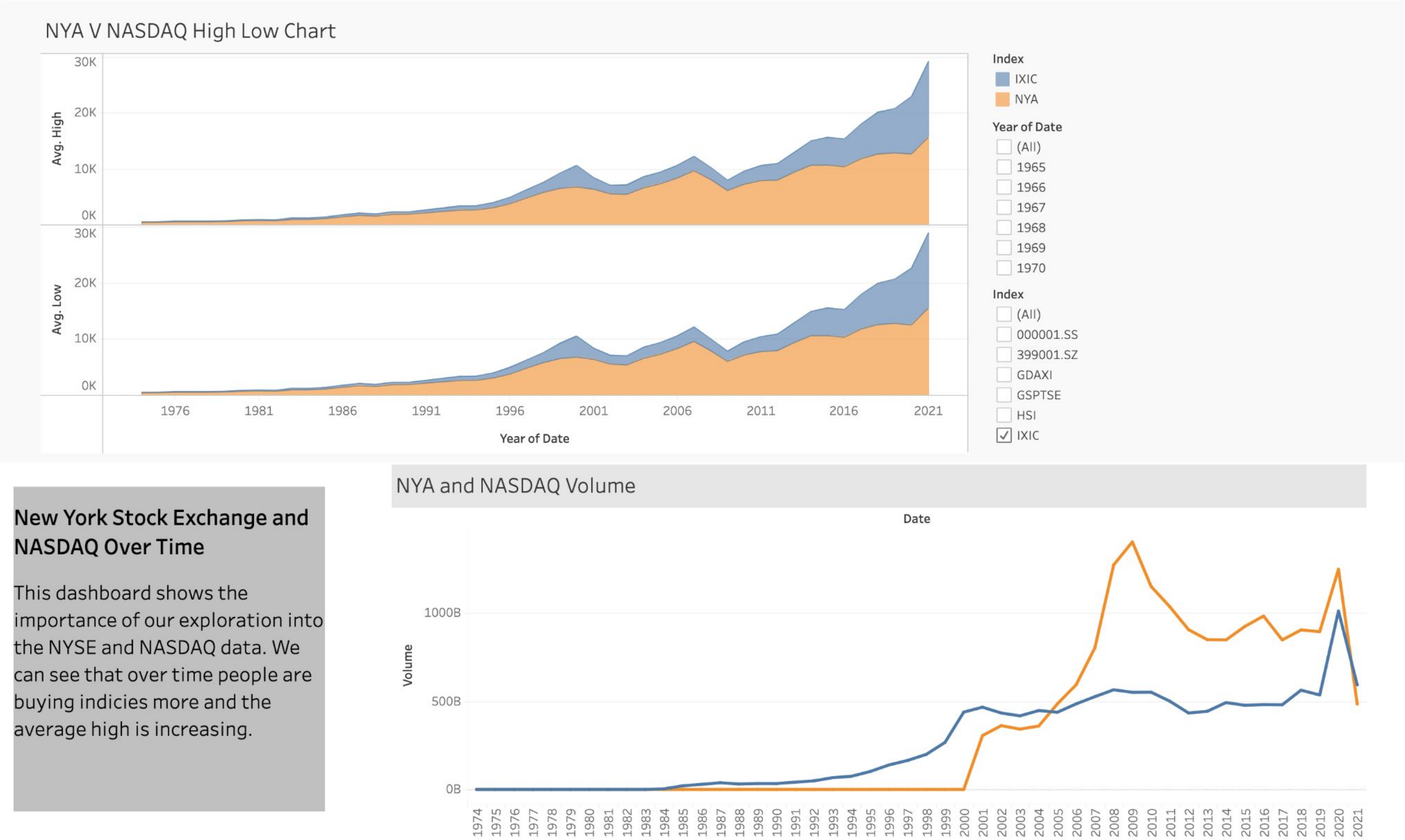
These charts show the overall distribution of Up Vs. Down since 1966


NYSE and NASDAQ - Up/Down from 1966-2021


NASDAQ - Up/Down from 1966-2021


NYSE - Up/Down from 1966-2021

EXPLORATORY ANALYSIS



NYA V NASDAQ High Low Chart

Index

🟦 IXIC

🟧 NYA

Year of Date

☐ (All)
☐ 1965
☐ 1966
☐ 1967
☐ 1968
☐ 1969
☐ 1970

Index

☐ (All)
☐ 000001.SS
☐ 399001.SZ
☐ GDAXI
☐ GSPTSE
☐ HSI
☑ IXIC

New York Stock Exchange and NASDAQ Over Time

This dashboard shows the importance of our exploration into the NYSE and NASDAQ data. We can see that over time people are buying indicies more and the average high is increasing.

NYA and NASDAQ Volume

ER Diagram

We started with 2 tables from Kaggle:

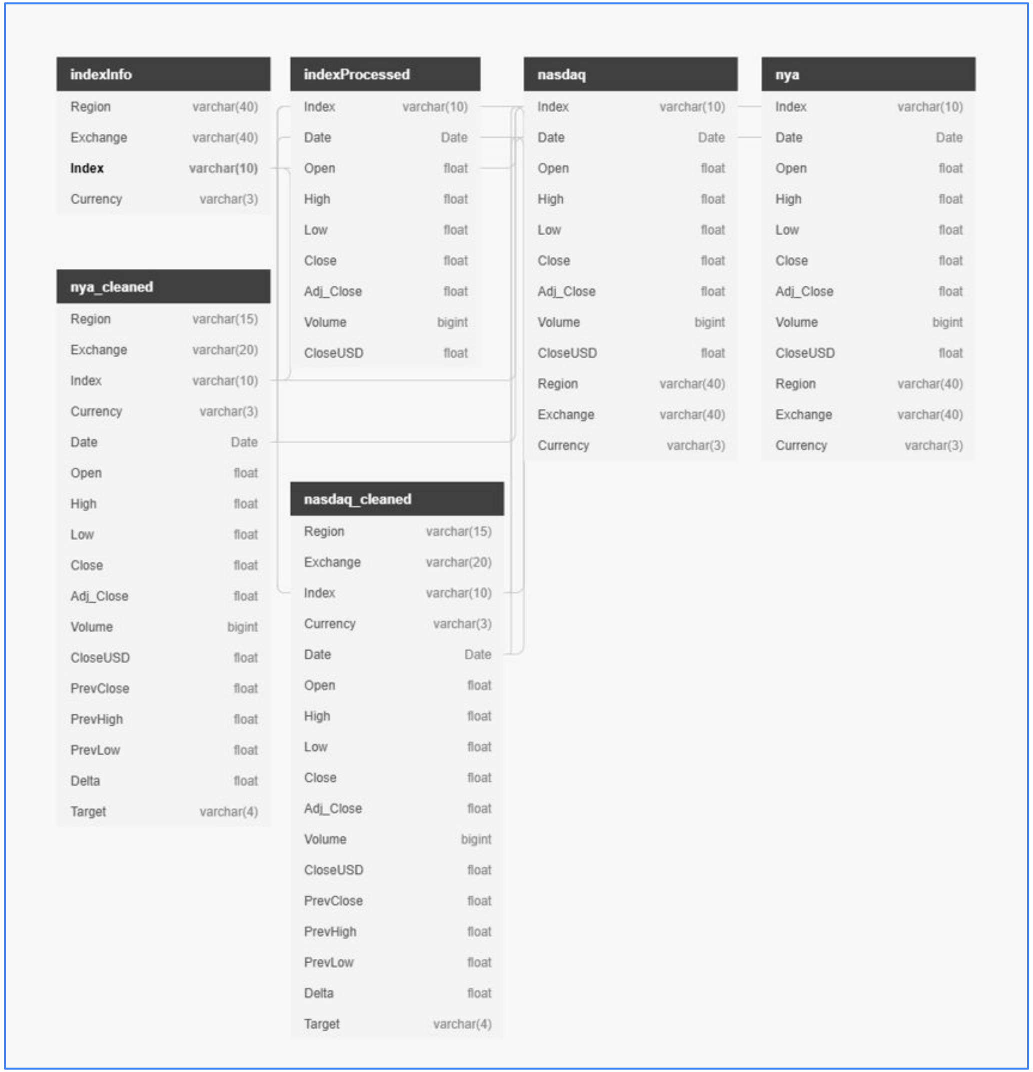
- indexInfo
- indexProcessed

We created 2 tables:

- Nasdaq & NYA - This was done by using a left Join on the “indexProcessed” and “indexInfo” table with a filter by index

We exported 2 tables back to SQL:

- NASDAQ_cleaned & NYA_cleaned - These tables are copied of the Nasdaq and NYA tables with new column names for data cleansing purposes



indexInfo	
Region	varchar(40)
Exchange	varchar(40)
Index	**varchar(10)**
Currency	varchar(3)

indexProcessed	
Index	varchar(10)
Date	Date
Open	float
High	float
Low	float
Close	float
Adj_Close	float
Volume	bigint
CloseUSD	float

nasdaq	
Index	varchar(10)
Date	Date
Open	float
High	float
Low	float
Close	float
Adj_Close	float
Volume	bigint
CloseUSD	float
Region	varchar(40)
Exchange	varchar(40)
Currency	varchar(3)

nya	
Index	varchar(10)
Date	Date
Open	float
High	float
Low	float
Close	float
Adj_Close	float
Volume	bigint
CloseUSD	float
Region	varchar(40)
Exchange	varchar(40)
Currency	varchar(3)

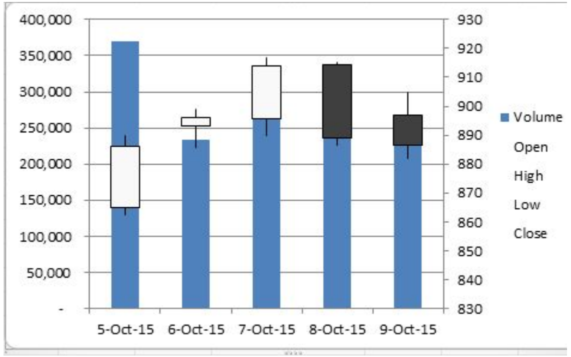
nya_cleaned	
Region	varchar(15)
Exchange	varchar(20)
Index	varchar(10)
Currency	varchar(3)
Date	Date
Open	float
High	float
Low	float
Close	float
Adj_Close	float
Volume	bigint
CloseUSD	float
PrevClose	float
PrevHigh	float
PrevLow	float
Delta	float
Target	varchar(4)

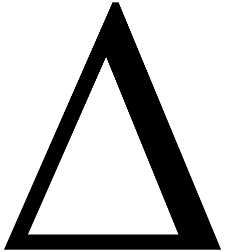
nasdaq_cleaned	
Region	varchar(15)
Exchange	varchar(20)
Index	varchar(10)
Currency	varchar(3)
Date	Date
Open	float
High	float
Low	float
Close	float
Adj_Close	float
Volume	bigint
CloseUSD	float
PrevClose	float
PrevHigh	float
PrevLow	float
Delta	float
Target	varchar(4)

VARIABLES

Name	Type	Explanation	NYA	NASDAQ
Index	Object	Index name (Eg: NYA, IXIC)	✓	✓
Date	Datetime	Date in question	✓	✓
Open	Float64	Opening value of the Index on the date	✓	✓
High	Float64	High value of the Index on the date	✓	✓
Low	Float64	Low value of the Index on the date	✓	✓
Close	Float64	Closing value of the Index on the date	✓	✓
Adj Close	Float64	Adjusted Closing value of the Index on the date	✓	✓
Volume	Int64	Volume traded of the Index on the date	✓	✓
CloseUSD	Float64	Closing value of the Index on the date in USD	✓	✓
Region	Object	Region where Index is located	✓	✓
Exchange	Object	Full name (Eg: New York Stock Exchange)	✓	✓
Currency	Object	Currency index is traded in	✓	✓

Data Cleansing & Wrangling





Dropped NaN/ Nulls

Added previous day's High, Low, Close to today's row

Added Delta and Target Columns

- Delta: % gain or loss based on Close price
- Target: Determine Up/Down Trend based on Delta

Machine Learning Models

Supervised Machine Learning Models Used:

- Logistic Regression
- Balanced Random Forest Classifier
- SMOTE oversampling
- Undersampling
- Decision Tree Model

Machine Learning Preparation:

We decided on the following **Features** for the model:

- Open
- PrevHigh
- PrevLow
- PrevClose

We tried to predict the **Target**: whether or not the index closed higher or lower than the previous day.

Logistic Regression

NYSE

Accuracy: 0.85

NASDAQ

Accuracy: 0.62

Drawback: Assumes linearity between dependent and independent variables

Success: Solid baseline!

Decision Tree Model

NYSE

Accuracy: 0.76
Precision: 0.76
F1: 0.76

NASDAQ

Accuracy: 0.63
Precision: 0.63
F1: 0.63

Drawback: Sensitive to small changes in data

Success: Visually intuitive and efficient!

Balanced Random Forest Classifier

NYSE

Accuracy: 0.77
Precision: 0.77
F1: 0.77

NASDAQ

Accuracy: 0.64
Precision: 0.64
F1: 0.64

Drawback: Value gained from additional samples drops off over time

Success: Handles linear and non-linear relationships well!

SMOTE Oversampling

NYSE

Accuracy: 0.87
Precision: 0.88
F1: 0.87

NASDAQ

Accuracy: 0.78
Precision: 0.78
F1: 0.78

Drawback: Overfitting more likely

Success: Doesn’t lose any information!

Undersampling

NYSE

Accuracy: 0.87
Precision: 0.87
F1: 0.87

NASDAQ

Accuracy: 0.77
Precision: 0.78
F1: 0.78

Drawback: Potentially discarding useful data

Success: Helps against skewing towards the majority class!

Machine Learning Model Results and Score Trends

Tableau

Results

It was better at predicting for the NYSE than NASDAQ

SMOTE Oversampling and Undersampling had the highest accuracy, precision, and F1 scores

The model was more accurate at predicting **higher** closings than it was at predicting **lower** closings

Accuracy was better when resampled

What We Would Do Differently

1

Zoom in on a particular stock or industry to see if there will be an accuracy improvement

2

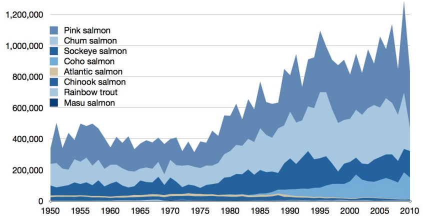
Look up existing machine learning progress/accuracy with stock predictions

3

Use a different model such as one that predicts continuous variables

Recommendations for Future Analysis

Predict closing *price*



Account for time series

HKEX 香港交易所

EURONEXT

Run model with different stock exchanges

Goldman Sachs

Predict closing direction for *individual* stocks

?