DILLON FARBER
CHRISTIAN DAVIS
QUINN TEMPLETON

PREDICTING THE STOCK MARKET

TABLE OF CONTENTS

DATASET INFORMATION

- The data was taken from the Kaggle datasets catalog.
- The dataset was comprised of multiple years of past stock information. 2014-2018.
- The intent of the dataset is to determine if the user should buy or sell the current stock they may be looking at.
- The dataset has over 20k rows with 225 columns

SUMMARY OF EXPERIMENT/QUESTION

The objective of this notebook is to take the stock information about a single stock and its information to find out whether the stock should be sold or bought. We take the information and use binary classification techniques to determine what the stock might be classified as to better determine the decision of someone to buy or sell that stock in its current state. We are using collected stock information from the years 2014-2018 years to train and test our models. The data has 20 thousand rows and 225 columns, giving us around 4 billion float and integer datapoints of data for us to use in creating models and feature selection.

TEXT

COLUMNS

stock
Revenue
Revenue Growth
Cost of Revenue
Gross Profit
R&D Expenses
SG&A Expense
Operating Expenses
Operating Income
Interest Expense
Earnings before Tax
Income Tax Expense
Net Income - Non-Controlling int
Net Income - Discontinued ops
Net Income
Preferred Dividends
Net Income Com
EPS
EPS Diluted
Weighted Average Shs Out
Weighted Average Shs Out (Dil)
Dividend per Share
Gross Margin
EBITDA Margin
EBIT Margin
Profit Margin
Free Cash Flow margin
EBITDA
EBIT
Consolidated Income
Earnings Before Tax Margin
Net Profit Margin
Cash and cash equivalents
Short-term investments
Cash and short-term investments
Receivables
Inventories
Total current assets
Property, Plant & Equipment Net

Income Quality
Dividend Yield
Payout Ratio
SG&A to Revenue
R&D to Revenue
Intangibles to Total Assets
Capex to Operating Cash Flow
Capex to Revenue
Capex to Depreciation
Stock-based compensation to Revenue
Graham Number
ROIC
Return on Tangible Assets
Graham Net-Net
Working Capital
Tangible Asset Value
Net Current Asset Value
Invested Capital
Average Receivables
Average Payables
Average Inventory
Days Sales Outstanding
Days Payables Outstanding
Days of Inventory on Hand
Receivables Turnover
Payables Turnover
Inventory Turnover
ROE
Capex per Share
Gross Profit Growth
EBIT Growth
Operating Income Growth
Net Income Growth
EPS Growth
EPS Diluted Growth
Weighted Average Shares Growth
Weighted Average Shares Diluted Growth
Dividends per Share Growth
Operating Cash Flow growth
Free Cash Flow growth
10Y Revenue Growth (per Share)

Long-term investments
Tax assets
Total non-current assets
Total assets
Payables
Short-term debt
Total current liabilities
Long-term debt
Total debt
Deferred revenue
Tax Liabilities
Deposit Liabilities
Total non-current liabilities
Total liabilities
Other comprehensive income
Retained earnings (deficit)
Total shareholders equity
Investments
Net Debt
Other Assets
Other Liabilities
Depreciation & Amortization
Stock-based compensation
Operating Cash Flow
Capital Expenditure
Acquisitions and disposals
Investment purchases and sales
Investing Cash flow
Issuance (repayment) of debt
Issuance (buybacks) of shares
Dividend payments
Financing Cash Flow
Effect of forex changes on cash
Net cash flow / Change in cash
Free Cash Flow
Net Cash/Marketcap
priceBookValueRatio
priceToBookRatio
priceToSalesRatio
priceEarningsRatio

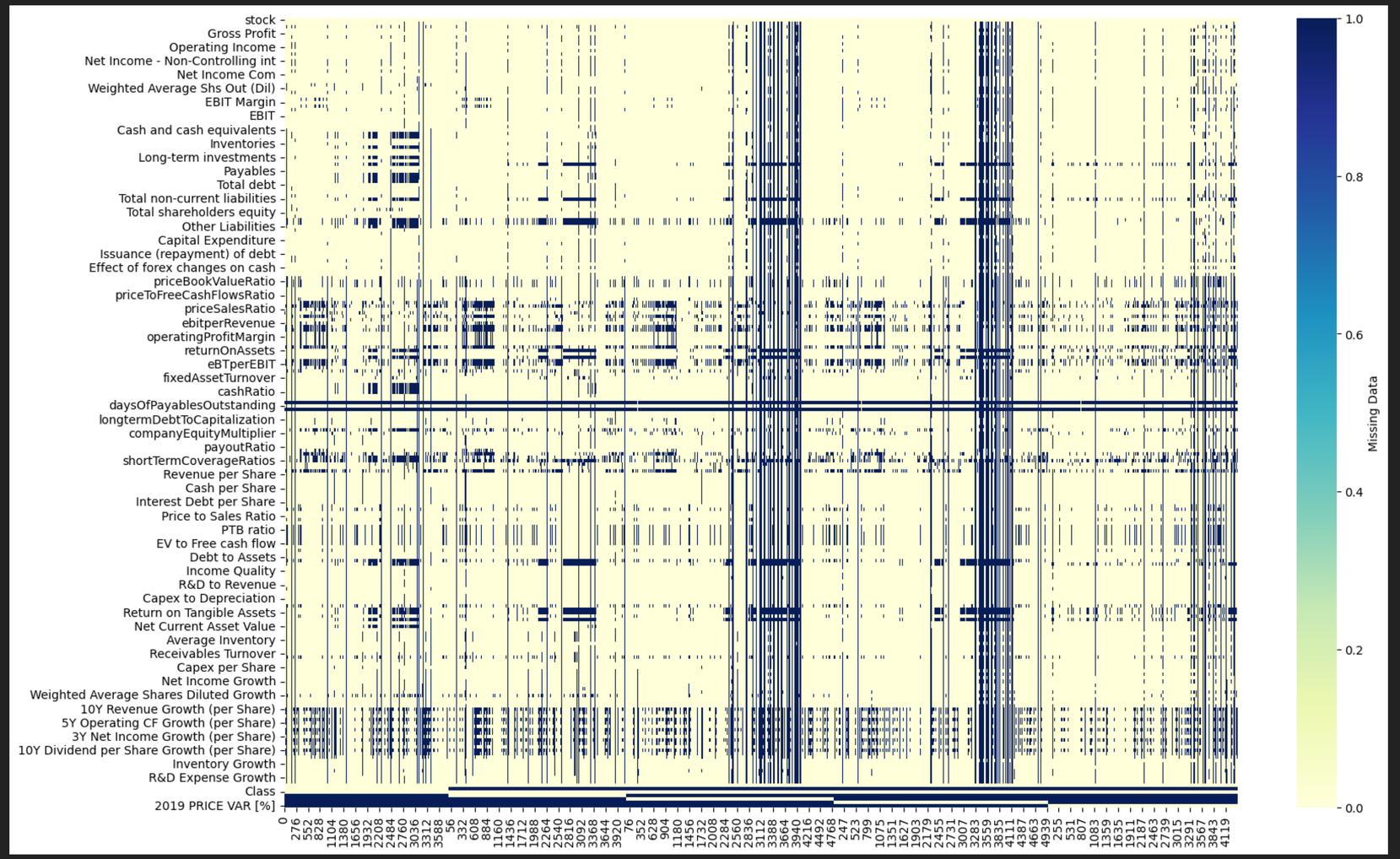
priceToFreeCashFlowsRatio
priceToOperatingCashFlowsRatio
priceCashFlowRatio
priceEarningsToGrowthRatio
priceSalesRatio
dividendYield
enterpriseValueMultiple
priceFairValue
ebitperRevenue
ebtperEBIT
niperEBT
grossProfitMargin
operatingProfitMargin
pretaxProfitMargin
netProfitMargin
effectiveTaxRate
returnOnAssets
returnOnEquity
returnOnCapitalEmployed
nIperEBT
eBTperEBIT
eBITperRevenue
payablesTurnover
inventoryTurnover
fixedAssetTurnover
assetTurnover
currentRatio
quickRatio
cashRatio
daysOfSalesOutstanding
daysOfInventoryOutstanding
operatingCycle
daysOfPayablesOutstanding
cashConversionCycle
debtRatio
debtEquityRatio
longtermDebtToCapitalization
totalDebtToCapitalization
interestCoverage
cashFlowToDebtRatio

companyEquityMultiplier
operatingCashFlowPerShare
freeCashFlowPerShare
cashPerShare
payoutRatio
operatingCashFlowSalesRatio
freeCashFlowOperatingCashFlowRatio
cashFlowCoverageRatios
shortTermCoverageRatios
capitalExpenditureCoverageRatios
dividendpaidAndCapexCoverageRatios
dividendPayoutRatio
Revenue per Share
Net Income per Share
Operating Cash Flow per Share
Free Cash Flow per Share
Cash per Share
Book Value per Share
Tangible Book Value per Share
Shareholders Equity per Share
Interest Debt per Share
Market Cap
Enterprise Value
PE ratio
Price to Sales Ratio
POCF ratio
PFCF ratio
PB ratio
PTB ratio
EV to Sales
Enterprise Value over EBITDA
EV to Operating cash flow
EV to Free cash flow
Earnings Yield
Free Cash Flow Yield
Debt to Equity
Debt to Assets
Net Debt to EBITDA
Current ratio
Interest Coverage

5Y Revenue Growth (per
3Y Revenue Growth (per
10Y Operating CF Growth (per
5Y Operating CF Growth (per
3Y Operating CF Growth (per
10Y Net Income Growth (per
5Y Net Income Growth (per
3Y Net Income Growth (per
10Y Shareholders Equity Growth (per
5Y Shareholders Equity Growth (per
3Y Shareholders Equity Growth (per
10Y Dividend per Share Growth (per
5Y Dividend per Share Growth (per
3Y Dividend per Share Growth (per
Receivables g
Inventory G
Asset G
Book Value per Share G
Debt G
R&D Expense G
SG&A Expenses G
pric

ZERO HEAT MAP

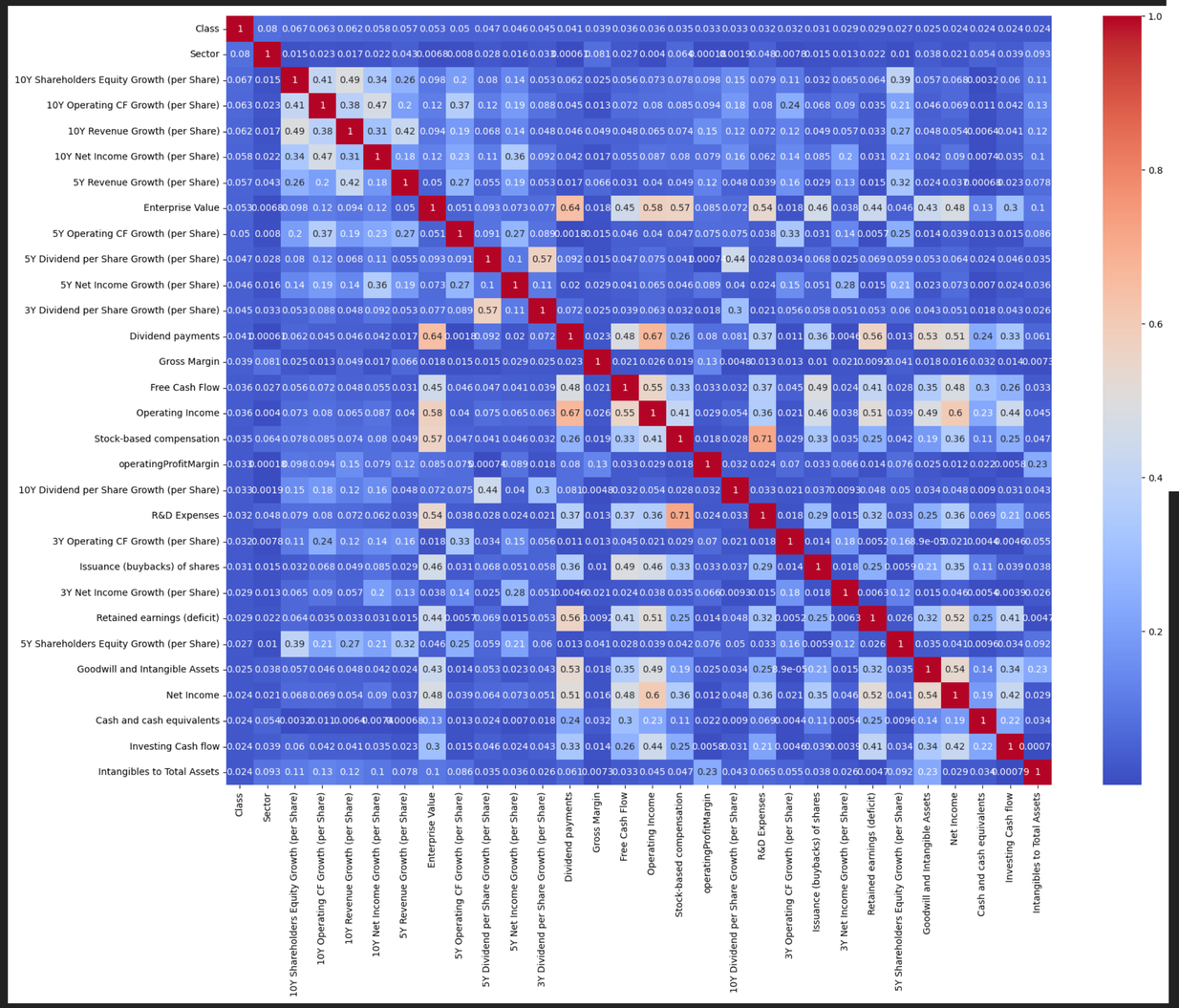
- We displayed visually to see what our data looked like and how it might affect the data by placing a zero in the columns.
- We did this because of the large number of missing values in the dataset.

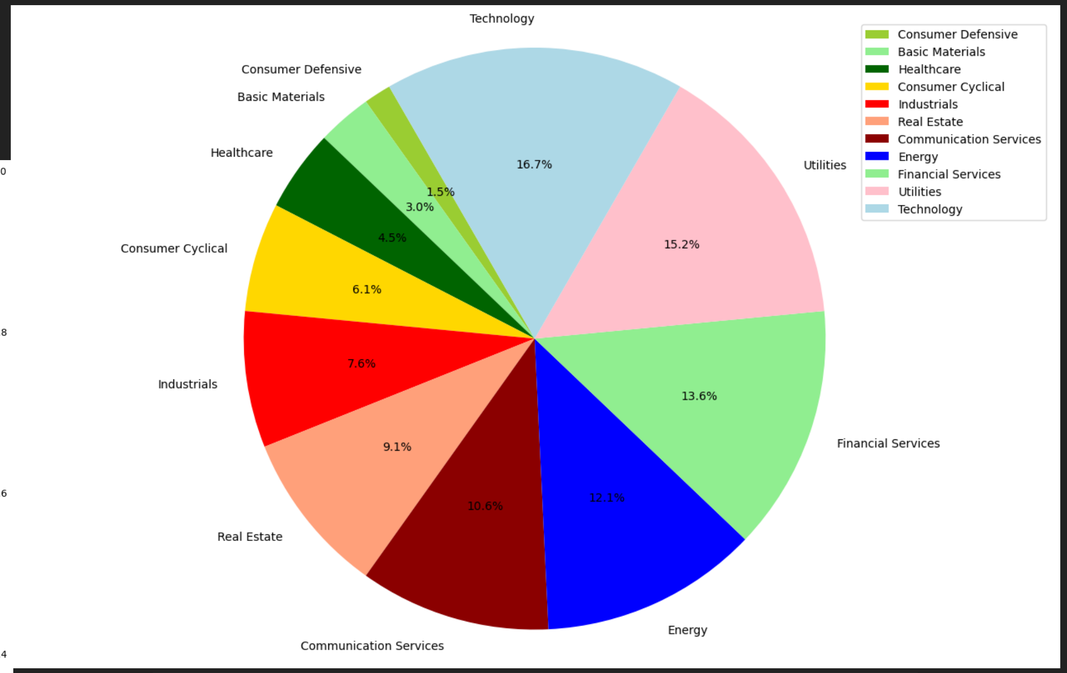


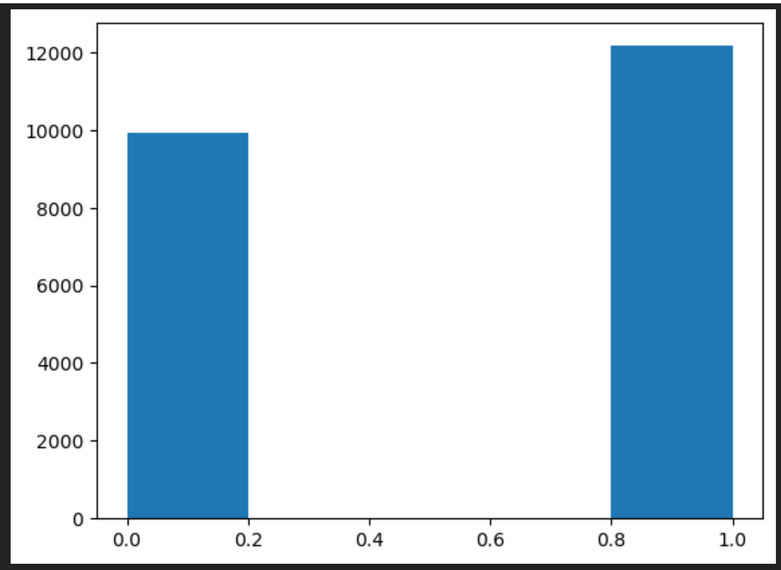
TRANSFORMING THE DATA

- Changed column names from text to integer values based on unique id.
- Filled missing data with 0’s to keep the majority of data, doing so left us with more information to use, doing a drop column left us with barely any data.
- Correlation map between columns and the column for our prediction class and looking for to highly correlated columns leading to duplicate columns.
- Bar chart for visual representation of the stability of our prediction column.

CORRELATION MAP/PIE CHART



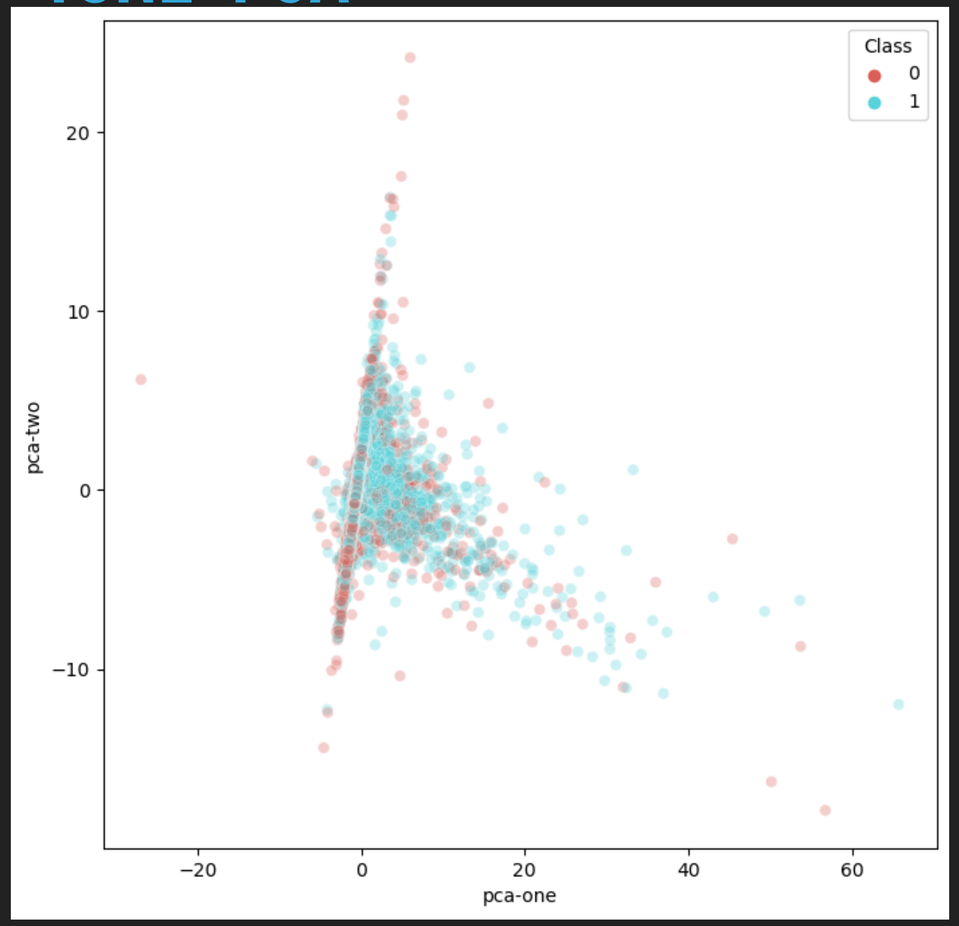




SCALING TECHNIQUE

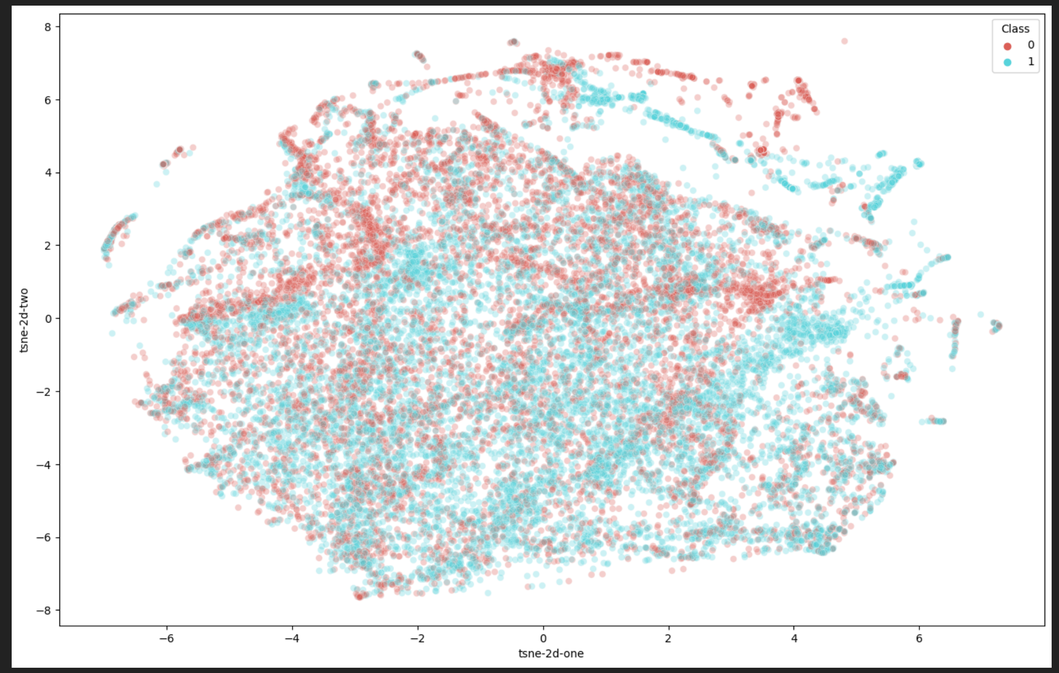
- Used Both Standard scaler and and the MaxAbsScaler().
- The Data had huge min and max differences, with minimum being in the negatives, and the max values in the 10’s of billions. With a mix in between the two.
- The Standard Scalar gave us better results, so we used this technique throughout the rest of the experiment.

TSNE-PCA



- PCA- used for high dimensionality, not has good as TSNE and used for linear dimensionality.

- We chose to use a dimensionality reduction tool and used both the TSNE and PCA to try and give our results the best chance.



- TSNE- used for high dimensionality in data, we received the best results from the set of data points. This reduction used gaussian when trying to reduce the distance between two points.

FEATURE SELECTION

- Several methods were used for feature selection.
  - Highest correlated columns
    - Took the 15 highest correlated features and used them as our data points.
  - Select K Best
    - Takes the best K value and then determines which column has the best score and then picks between them and returns those columns back as our x.
  - Select Percentile
    - This takes in a function for scoring each column, Mutual info classifier, then it selects the top percentile that you chose.

GRADIENT BOOSTING

- Takes a data set and puts it through training on a tree.
- Then it takes the predictions and uses the residual errors and then trains the dataset.
- The model keeps doing this till it each tree is trained.
- The model is then ready for prediction testing.

K NEAREST NEIGHBOR

- Supervised learning technique.
- The model is trained by taking the n-nearest neighbors and predicts what the outcome might be.
- The model groups data and then uses these groups to determine the nearest neighbors to the dataset to be predicted.
- The algorithm returns the classification of the group of n-nearest neighbors.

LOGISTIC REGRESSION

- Takes multiple labels in and the predictor is a binary operation
- The model makes a threshold based on one or multiple labels give as x.
- Once the threshold is in place, it takes that and predicts the outcome.
- If below its 0 and if above it is 1.

SVM

- Using the SVM in sklearn with the SVC(Support Vector Clustering) model.
- It makes a line that then gives a support vectors to the nearest x to the decision line.
- If above and below the line and outside the margin of error we classify as one or the other, but in side the the margins or support vectors the model uses the decision line to make the determination.

CONCLUSION/RESULTS

- The outcome, we see that KNN was our best model, closely followed by Gradient boosting. The SVC model was our worst model for our data . We were able to get a 78% accuracy rate on our data with KNN with the TSNE technique. This is a pretty good score for predicting the buy or sell of a stock.