Riya Senthil

110 Galway Cir, Chalfont, PA 18914
riyasenthil8@gmail.com

EDUCATION

Central Bucks High School South, Warrington, PA **Expected Graduation: June 2022**

- AP Coursework
  - Completed: Chemistry (5), Biology (4), Computer Science Principles(4), Calculus AB (5).
  - Ongoing: Calculus BC, Physics (Newtonian Mechanics), Statistics, Macroeconomics.
- GPA: Unweighted: 3.7949 Weighted: 4.2179
- SAT Super score: 1490. Math: 780, ERW: 710

Summer Institute for the Gifted (SIG) at Princeton University **Summer 2018**

- Coursework in engineering, genetics, algebra, and microbiology

University of Pennsylvania Girls in Engineering Math and Science Program (Penn GEMS) **Summer 2018**

- One of 70 students accepted from a pool of 300 applicants

HONORS & AWARDS

2022 Commended National Merit Scholar

- Scored within top 50,000 students nationwide who took the PSAT in 2020

Johns Hopkins Center for Talented Youth Advanced Level in English and Math

First Lego League (FLL) 3rd Place Robotics Award

- Won 3rd place out of 45 regional finalists in UPenn at 2018 Southeast Finals

Shaolin Kempo Karate 2nd Degree Black Belt

- 8 years of training and knowledge of over 100 combinations and 14 forms

2019 1st Place Sparring, Villari's Shaolin Kempo Karate PA Regional Tournament

2018 2nd Place Sparring, Villari's Shaolin Kempo Karate PA Regional Tournament

2018 4th Place Self-Defense, Villari's Shaolin Kempo Karate PA Regional Tournament

2018 Robert M. Gendall Award: given to 4 students out of 270 in my grade for exemplary class participation

EXPERIENCE

Two Research Projects: (1) Predicting Recessions and (2) Predicting Breast Cancer using Machine Learning **2021**

- Guided by Prof. Badrinath Kottimukkalur of George Washington University
- Predicted the presence of malignant breast cancer cells using features from histological cell images.
- Predicted Recessions using data from Federal Reserve Economic Data
- Self-taught Python coding language

Volunteer at Lansdale Abington Jefferson Hospital **March 2020-present**

- Manage front desk duties, such as checking patients in, delivering packages, and communicating patient statuses with nurses
- 4-5 hours a week, 47 weeks a year.

Blood Drive Coordinator, Red Cross Club **December 2019-present**

- Coordinated blood drive of over 60 people
- Raised $160 for the Red Cross

FLL Robotics Junior Mentor **Present**

- Taught children robotics building and coding

FLL Robotics Leader and Member **2018-2019**

- Founded all-female robotics team, recruited 4 peers, and participated in tournaments.
- Designed, developed and coded multiple robots to complete competitive tasks

Leadership Program at Villari's Self-Defense **December 2018-present**

- Organize and set up tournaments
- Volunteer to help younger kids with their forms
- Organize fundraisers for Manna on Main Street, raising ~$11,000 each year

CLUBS AND ACTIVITIES

Villari's Shaolin Kempo Karate **2013-Present**

- Second-degree black belt

First Tech Challenge: Robotics **October 2021-Present**

- Member of school team to participate in robotics competitions

National Honors Society **October 2020-Present**

Science National Honor Society **October 2020-Present**

Titan Senate (Student Council) **September 2020-Present**

- Organized set and created COVID precaution rules for Junior Prom

Young Democratic Socialists of America (YDSA) **September 2020-Present**

- Advocate for social issues such as walking door to door to encourage people to vote in the school board election
- Raised money for a bake sale for LGBTQ+ event, raising a total of $300 dollars

Chem Club **September 2019-Present**

Health Occupations Students Of America (HOSA) **September 2019-Present**

- Learned about professions in the health field
- Designed Healthy Lifestyle projects

First Lego League (FLL) Robotics **2017-2019**

- Dedicated 5-6 hours a week to robotics
- Designed, developed, and coded a robot
- Won overall 1st place award and 1st place robotics award at Hatboro Horsham Qualifier Event in 2017 and 2018
- Won Teamwork award and 3rd place robotics award at UPenn regionals in 2019

Red Cross Club **September 2019-Present**

- Help in my community by raising money for the red cross
- Spent a total of 25+ hours volunteering

Four Diamond Minithon **November 2019**

- 6 hour event dedicated to fighting childhood cancer
- My group raised a total of $200 for cancer

Breast Cancer Diagnosis Using Machine Learning

Riya Senthil

Guided By: Dr. Badrinath Kottimukkular, Assistant Professor, George Washington University, Washington DC.

Objective

Using digitally coded histopathological image data, predict the samples that has breast cancer (malignant) cells. Identify which features in the images are highly predictive of malignancy.

Data Description

The breast cancer diagnosis data is a publicly available dataset [1]. There are 699 observations, 9 features and diagnosis indicating the breast cancer cells are malignant or benign.

After a Fine Needle Aspiration (FNA) biopsy, microscopic examination produces histopathological images of the cells. These cells are then coded into various characteristics and the breast cancer dataset is generated.

The nine features are all coded between 1 to 10.

Features are:

- clump thickness – malignant sample has thick grouping of cancer cells in multiple layers.
- uniformity of cell size – represents metastasis to lymph nodes.
- uniformity of cell shapes - cancerous cells have varying sizes.
- marginal adhesion - suggests loss of adhesion which is a sign of malignancy.
- single epithelial cell size (SECS) – Larger SECS may indicate a malignant cell.
- bare nuclei – benign cells has a smaller number of nuclei without cytoplasm. Sixteen missing values.
- bland chromatin - In cancer cells the chromatin tends to have coarse texture.
- normal nucleoli - generally very small in benign cells.
- mitoses – cell division process. Generally higher counts in malignant cells.

	Clump_Thickness	Cell_Size_Uniformity	Cell_Shape_Uniformity	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Diagnosis
count	699	699	699	699	699	683	699	699	699	699
mean	4.42	3.13	3.21	2.81	3.22	3.54	3.44	2.87	1.59	0.34
std	2.82	3.05	2.97	2.86	2.21	3.64	2.44	3.05	1.72	0.48
median	4	1	1	1	2	1	3	1	1	0

Table 1. Descriptive feature statistics. Bare Nuclei has 683 non-missing values. 34% of samples have malignant diagnosis.

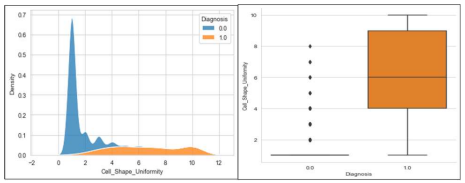

Fig. 1 Left chart shows Clump Shape Uniformity density distributions of benign (blue, diagnosis = 0) and malignant cells (orange, diagnosis = 1). Right chart shows same distributions in a box plot. There is a strong differentiation in values of Clump Shape Uniformity between benign and malignant cells.

Feature Exploration

All nine features are converted to float. The target variable Diagnosis is converted to 0 and 1 representing benign and malignant sample, respectively. Python packages such as *numpy* and *pandas* are used for computations. Packages *matplotlib* and *seaborn* are used for plotting and data visualization.

Table 1 has descriptive statistics. There are 699 observations. Bare Nuclei has 16 missing values. Mitoses, Normal Nuclei and Marginal Adhesion have most records with low values (avg: 1.6 to 2.8). All other features also have low values in general (avg: 3.1 to 4.4). In most cases, standard deviations are similar or less than mean. About 34% of observations are malignant samples and 66% are benign samples. Data is a bit unbalanced with more benign samples.

Figure 1 shows that cell shape uniformity value is low (i.e., uniformly shaped cells) in benign cells. Cancer cells have high cell shape uniformity value (irregular cell shapes) and is well differentiated from normal cells. Thus Cell Shape Uniformity is a good predictor of malignancy. Many features have similar strong differentiation between benign and malignant cells. Example: Cell Size Uniformity, Bare Nuclei, Normal Nucleoli, Single Epithelial Cell Size.

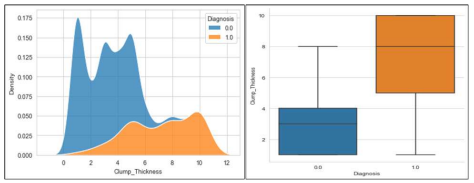

Fig. 2 Left chart shows Clump Thickness density distributions of benign (blue, diagnosis = 0) and malignant cells (orange, diagnosis = 1). Right chart shows same distribution in a box plot. There is some overlap and a medium level of differentiation in values of Clump Thickness between benign and malignant cells. Distributions deviates fairly from normal distribution.

Figure 2 shows clump thickness (grouping of cancer cells into multiple layers) is higher in malignant cells and is somewhat differentiated from benign cells. However, there is a fair overlap in clump thickness of benign and malignant cells. They are not normally distributed and has bands and tails. Few other variables show similar overlaps in their values. Example: Mitoses, Marginal Adhesion.

Many of the variables have moderate to high correlations among themselves. Highly correlated features are Cell Size, Shape and Texture uniformity (0.76 to 0.91). These uniformity values along with bare nuclei count have highest correlation (0.82) with benign vs malignant diagnosis. Mitoses has relatively lower correlation with other variables (0.34 to 0.48).

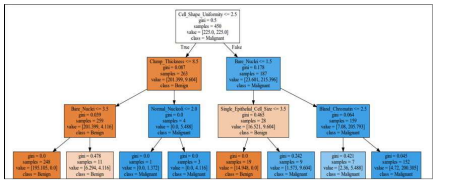

Fig. 3. Decision tree with max depth of three and using diagnosis as target. Orange nodes denote benign cell classifications, and the blue nodes denote malignant cell classification. Most important features to classify benign vs malignant cells are Cell Shape Uniformity, Clump Thickness, Bare Nuclei and Single Epithelial Cell Size.

Models and Results

Data is split into 66% for training [450 observations] and 34% for testing [283 observations]. Models are built on training test dataset and their performances are evaluated using test data. Least error in classifying the malignant cells is emphasized. Training data is balanced (i.e., malignant samples are repeated) so that the data has equal representation of benign and malignant cells to be modelled. Data balancing helps to improve model accuracy.

Decision trees classify target variable. The algorithm first goes through all the features and chooses a feature and boundary point that has maximum information to classify the target variable. Using this boundary, it subsets the data into two nodes. The algorithm is then iterated for each of the child node until leaf nodes with maximum classification accuracy could be obtained. Tree depth is adjusted to not overfit the data. Python packages sci-kit learn (*sklearn*) and *graphviz* were used in building decision trees. Figure 3 shows a tree of depth 3 with diagnoses as target variable and all nine features as inputs. Most important features to classify benign vs malignant cells are **Cell Shape Uniformity, Clump Thickness, Bare Nuclei and Single Epithelial Cell Size**.

Logistic regression classifies the binary outcomes (benign vs malignant) using regression methods that seek to fit a linear equation that minimizes the error in classification. A logit transformation is done to the output variable so that the output is the probability of a given sample to have a value of 1 (i.e., malignant). Odds ratio, that represents odds of a given feature to be predictive of malignancy, is used to choose most influential features to predict malignancy. Python package *statsmodels* was used for logistic regression models. Table 2 shows results from logistic regression model where some of the non-significant variables were removed. Higher odds ratios are seen *(1.5 to 1.8)* for ***Clump Thickness, Cell Shape Uniformity, Bare Nuclei and Bland Chromatin in predicting malignancy.***

	Coefficient	Standard Error [Std. Deviation of the coefficient]	Confidence Level	Odds Ratio
Intercept	-9.3061	1.079	100%	0.00
Clump_Thickness	0.5836	0.145	100%	1.79
Cell_Shape_Uniformity	0.405	0.184	97%	1.50
Marginal_Adhesion	0.2955	0.165	93%	1.34
Bare_Nuclei	0.4059	0.103	100%	1.50
Bland_Chromatin	0.3837	0.172	97%	1.47
Normal_Nucleoli	0.2109	0.108	95%	1.23
Mitoses	0.5369	0.36	87%	1.71

Table 2: Results from logistic regression with select features. Except for Mitoses and Marginal Adhesion all other features are statistically significant at 95% confidence level. Clump Thickness, Cell Shape Uniformity, Bare Nuclei and Bland Chromatin have higher odds (1.5 to 1.8) in predicting malignancy.

Confusion Matrix. (Cells have count of records)		Predicted: Negative [=0] (benign)	Predicted: Positive [=1] (malignant)	
Actual	Negative [=0] (benign)	True Negative [TN]	False Positive [FP]	Accuracy = [TP + TN] / [TP + TN + FP + FN]
	Positive [=1] (malignant)	False Negative [FN]	True Positive [TP]	Precision = [TP] / [TP + FP]
				Recall = [TP] / [TP + FN]
				F1 = [2*Precision*Recall] / [Precision + Recall]

Fig. 4: A confusion matrix is used to assess the model performance. It compares actual vs predicted classification record counts in test dataset. Various metrics such as accuracy, precision, recall and F1 are used to assess the model performance. Metric values range between 0 (low performance) to 1 (perfect prediction). Here we focus more on Recall as the cost of predicting someone has no cancer when they actually have cancer is high.

Best Model and Results

A confusion matrix (Fig. 4) produces various classification accuracies. **Accuracy** defines overall predictive accuracy as % of samples that were classified correctly as either benign or malignant. **Recall** represents % of actual malignant samples that were correctly predicted. **Precision** represents % of predicted malignant samples that are actually malignant. **F1 score** is a weighted average of Recall and Precision. **Best models are chosen by analyzing the highest recall in predicting malignancy using test dataset.**

Table 3 shows performance metrics for various model configurations. **Logistic regression with select variable (Table 2) is the best predictive model with a recall of 97.8% and accuracy of 97.4%. Decision Tree model with balanced data and a max depth of 3 (Fig. 4) has almost similar recall of 97.3% and is straight forward to interpret.** We use this decision tree model to explain the results and are discussed in conclusion section.

Model Description		Recall	F1	Precision	Accuracy
Decision Trees	Unbalanced data. Unrestricted tree depth (=9)	0.93	0.93	0.94	0.96
	Balanced data. Unrestricted tree depth (=11)	0.91	0.925	0.94	0.95
	Balanced data with depth 3	**0.973**	0.948	0.924	0.965
	Balanced data with depth 5	0.933	0.933	0.933	0.957
Logistic Regression	All Features	0.978	0.967	0.957	0.974
	Non significant features removed	**0.978**	0.967	0.957	0.974

Table 3: Classification metrics from confusion matrix for various model configurations. Logistic regressions with statistically non-significant features removed has the highest recall at 97.8% and a precision of 97.4% . This is the best predictive model. It is closely followed by random forest model and decision tree with max depth of 3, with recalls at 97.3% and predictive accuracy of 96.5%. Decision tree models are more straight forward to interpret.

Conclusions

The breast cancer dataset [1,2] with 9 digitally coded features from histopathological images of the breast cancer cell samples is used to predict if the given sample is benign or malignant. This data was split into 66% for training and 34% for testing. Classification models such as decision trees and logistic regression were applied to training data with diagnosis as target variable. Logistic regression with select variables with highest recall of 97.8% is the best predictive model followed closely by a decision tree model with short tree depth and a recall of 97.3%. Based on this easily interpretable decision tree model following was observed:

A given sample is malignant when:

- Cell Shape is more uniform (Cell_Shape_Uniformity <= 2.5) BUT the Clump Thickness is very high (Clump_Thickness > 8.5)
- Cell shape is not so uniform (Cell_Shape_Uniformity > 2.5) AND there are more bare nuclei (Bare_Nuclei > 1.5) AND a coarse texture of chromatin (Bland_Chromatin > 2.5). Even if chromatin has a fairly uniform texture, there is a high chance of malignancy.
- Cell shape is not so uniform (Cell_Shape_Uniformity > 2.5) AND there are less bare nuclei (Bare_Nuclei <= 1.5) BUT the single epithelial cell size is large (Single_Epithelial_Cell_Size > 3.5). In this case, there is a fairly good chance of malignancy.

References

[1] Data Source: http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html
Dr. WIlliam H. Wolberg , University of Wisconsin Hospitals, Madison, Wisconsin, USA
Donor: Olvi Mangasarian (mangasarian@cs.wisc.edu). Received by David W. Aha

[2] W.H. Wolberg, W.N. Street, and O.L. Mangasarian.
Breast Cytology Diagnosis via Digital Image Analysis. Analytical and Quantitative Cytology and Histology, Vol. 15 No. 6, pages 396-404, December 1993.

Predicting Recession From Economic Indicators

Riya Senthil

Guided By: Dr. Badrinath Kottimukkular, Assistant Professor, George Washington University, Washington DC.

Objective

Using quarterly economic indicators from Federal Reserve database, predict if a given quarter is under recession. Identify the key economic indicators that are highly predictive of recession.

Data Description

The Federal Reserve Economic Data is a publicly available dataset [1]. The data source has downloadable economic indicators aggregated daily or monthly or quarterly. During data preparation, these features are aggregated at quarterly level and 241 quarters that has most non-missing feature values from Q1 1961 to Q1 2021 is used in the analysis. Some variables with a lot of missing values are later dropped. Additional features are derived from original data to better represent the market reality in predicting recession.

Target variable and Features are:

- Target variable: Quarterly Recession Indicator (USRECQ). Recession:1 No-Recession: 0.
- Change in real private inventories (CBIC1). Represents inventory volume change from a year ago quarter.
- Permits issued for new privately-owned housing units in the quarter (PERMIT_SUM)
- Derived: Maximum monthly change in housing permits within the quarter (PERMIT_DIFF)
- Volume of retail trade sales in the quarter (SLR). Data missing from Q3 2018.
- Derived: Year over Year percentage change in retail trade sales (YoYC_PER_SLR)
- Average quarterly change in treasury bill yield between 10 and 2 years (T10_Mean).
- Average unemployment rate in each quarter (UNRATE_MEAN)
- Derived: Maximum monthly change in unemployment rate within the quarter (UNRATE_DIFF).
- Housing price index in each quarter (HPI)
- Derived: Year over Year percentage change in housing price index (YoYC_PER_HPI)
- Derived Lagged Features (lg1_*): All the above features were also lagged by one quarter. These are used to predict recession in each quarter using economic indicators from previous quarter.

	USRECQ	CBIC1	PERMIT_SUM	SLR	YoYC_PER_SLR	T10_MEAN	UNRATE_MEAN	HPI	YoYC_PER_HPI
count	241	241	241	230	230	19	241	185	181
mean	0.14	35.0	4098	68.3	2.03	0.59	6.01	235.4	4.7
std	0.34	57.4	1143	20.5	3.7	0.37	1.68	114.5	4.27
median	0	35.7	4051	64.6	2.61	0.51	5.7	208.0	5.08

Table 1. Descriptive statistics of various features. 241 quarters range from Q1 1961 to Q1 2021. Treasury yield (T10_MEAN) has only 19 quarters of data. Retail sales (SLR) and Housing Price Index (HPI) have data for 230 and 185 quarters respectively. Only 14% of quarters are in recession.

Feature Exploration

Python packages such as *numpy* and *pandas* are used for computations. Packages *matplotlib* and *seaborn* are used for data visualization. Table 1 has descriptive statistics. There are 241 quarters from Q1 2016 to Q1 2021. The change in treasury bill yield (T10_MEAN) has only 19 quarters worth of data starting from Q3 2016. Retail sales is missing last 11 quarters. Housing price Index has 56 missing quarters.

Feature ranges do vary in magnitude (ex: avg of 0.59 T10_MEAN to 4098 PERMIT_SUM). USRECQ, CBIC1, YoYC_PER_SLR and YoYC_PER_HPI have high variations with standard deviations higher than their means. About 14% of quarters have recession and 86% are no-recession quarters. Data is unbalanced and has more no-recession quarters.

Figure 1 shows that the YoY change in retail sales (YoY_PER_SLR) has distribution skewed slightly to the left during quarters of no-recession. During recession quarters, the data is fairly normally distributed. The Box plot shows that recession quarters have a higher decrease in year over year change in retail sales and its values has a good separation between no-recession and recession quarters. Similar visual inspection suggests following features are likely to be important predictors of recession: year over year change in retail sales (YoY_PER_SLR), change in inventories (CBIC1), unemployment rate (UNRATE_MEAN), year over year change in housing price index (YoYC_PER_HPI) and housing permits (PERMIT_SUM).

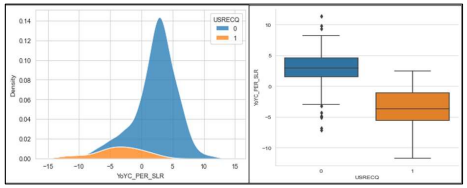

Fig. 1 Left chart shows year over year percent change in quarterly retail sales (YoY_PER_SLR) density distributions of no-recession (blue, USRECQ = 0) and recession quarters (orange, USRECQ = 1). Right chart shows same distribution in a box plot. There is a reasonably strong differentiation in values of YoY_PER_SLR between no-recession and recession quarters.

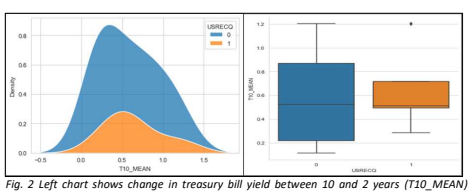

Fig. 2 Left chart shows change in treasury bill yield between 10 and 2 years (T10_MEAN) density distributions of no-recession (blue, USRECQ = 0) and recession quarters (orange, USRECQ = 1). Right chart shows same distribution in a box plot.

Figure 2 shows that the change in treasury bill yield between 10 and 2 years (T10_MEAN) has distribution skewed to the right.. The Box plot shows a big overlap between recession and no-recession, so this variable is not a good predicter. Since it also has lot of missing values, this feature is dropped from further analysis.

Recession is negatively correlated to YoY change in retail sales (-0.58) and change in inventory (-0.48). Other variables that are highly directly correlated among each other are housing permits issued and year over year change in housing price index (0.76).

Models and Results

Modeling data has 169 observations and 15 predictors with no missing data. Predictors include lagged and derived features. Data is split into 80% for training [n=135] and 20% for testing [n=34]. Training and Test data have 9.6% and 14.6% of recession quarters.

Models are built on training data and their performances are evaluated using test data. We emphasis on least error in classifying the recession quarters. Training data is balanced (i.e., recession quarters are repeated) to have equal representation of no-recession and recession quarters. Data balancing helps to improve model accuracy.

Decision trees are supervised learning methods. The algorithm first goes through the features and chooses a feature and boundary point that has maximum information to classify the target variable. Using this boundary, it subsets the data into true and false nodes. Process is repeated for each of the child nodes until leaf nodes with maximum classification accuracy could be obtained. Tree depth is adjusted to minimize overfitting.. Python packages sci-kit learn *(sklearn)* and *graphviz* were used in building decision trees. Figure 3 shows a tree of depth 4 and balanced data with USRECQ as target variable. Most important economic indicators to identify recession are Lagged YoY % change in retail sales (lg1_YoYC_PER_SLR), Change in unemployment rate within the quarter [UNRATE_DIFF], Lagged change in housing permits [lg1_PERMIT_DIFF] and Lagged YoY % change in housing price index [lg1_YoYC_PER_HPI].

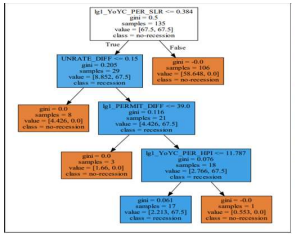

Fig. 3. Decision tree with balanced data, a max depth of four and using USRECQ as target. Orange nodes denote no-recession, and the blue nodes denote recession. Most important features to identify recession quarters are Lagged YoY % change in retail sales (lg1_YoYC_PER_SLR), Change in unemployment rate within the quarter [UNRATE_DIFF], Lagged change in housing permits [lg1_PERMIT_DIFF] and Lagged YoY % change in housing price index [lg1_YoYC_PER_HPI].

It is advantageous to predict recession in next quarter using economic indicators of current quarter. Such a decision tree model is built by using only the lagged feature values as predictor of current quarter recession. Predictive performance of this model is slightly below that of the previous model with all features.

Logistic regression classifies the binary outcomes (recession vs no-recession) using regression methods that seek to fit a linear equation that minimizes the error in classification. A logit transformation is done to the output variable so that the output is the probability of getting a value of 1 (i.e., recession quarter). We observed that the logistic regression was inferior in predicting recession.

Confusion Matrix. (Cells have count of records)		Predicted: Negative [=0] (benign)	Predicted: Positive [=1] (malignant)	
Actual	Negative [=0] (benign)	True Negative [TN]	False Positive [FP]	Accuracy = [TP + TN] / [TP + TN + FP + FN] Precision = [TP] / [TP + FP]
	Positive [=1] (malignant)	False Negative [FN]	True Positive [TP]	Recall = [TP] / [TP + FN] F1 = [2*Precision*Recall] / [Precision + Recall]

Fig. 4: A confusion matrix is used to assess the model performance. It compares actual vs predicted classification record counts in test dataset. Various metrics such as accuracy, precision, recall and F1 are used to assess the model performance. Metric values range between 0 (low performance) to 1 (perfect prediction). Here we focus more on Recall as the cost of predicting a quarter has no recession when it actually is in recession is high.

Best Model and Results

A confusion matrix (Fig. 4) produces various classification accuracies. **Accuracy** defines overall predictive accuracy as % of samples that were classified correctly as either recession or no-recession quarter. **Recall** represents % of actual recession quarters that were correctly predicted. **Precision** represents % of predicted recession quarters that actually has recession. **F1 score** is a weighted average of Recall and Precision. **Best models are chosen by analyzing the highest recall in predicting recession quarters using test dataset.**

Table 3 shows performance metrics for various model configurations. Logistic regression model has a poor performance with just 50% recall and precision. **Decision Trees [Fig. 3] using balanced data with a depth of 4 has the highest recall at 100% and a precision of 83%. This is the best predictive model and is easy to interpret.** We use this decision tree model to explain the results and are discussed in conclusion section.

	Model Description	Recall	F1	Precision	Accuracy
Decision Trees	Unbalanced data. Unrestricted tree depth (=5)	0.4	0.6	1	0.91
	Balanced data. Unrestricted tree depth (=8)	0.8	0.8	0.8	0.94
	Balanced data with depth 4	1	0.91	0.83	0.97
	Balanced data with depth 4 - Lagged Predictors only	0.88	0.82	0.77	0.93
Logistic Regression	Balanced data & Non significant features removed	0.5	0.5	0.5	0.88

Table 3: Classification metrics from confusion matrix for various model configurations. Decision Trees using balanced data with a depth of 4 has the highest recall at 100% and a precision of 83%. This is the best predictive model and is easy to interpret. Logistic regression model has a poor performance with just 50% recall and precision.

Conclusions

The Federal Reserve Economic Data [1] is used to predict if a given quarter is in recession or not. The data is processed to obtain quarter level aggregated features. Relevant features like percent changes and maximum within quarter changes were further derived. One quarter lagged features are also computed as additional features. Data is considered from Q1-1961 to Q1-2021. However, after removing missing values, there are 161 quarters to model. This data is split into 80% for training and 20% for testing. Classification models such as decision trees and logistic regression are applied to training data with recession indicator as target variable. Decision Trees using balanced data with a depth of 4 has the highest recall at 100% and a precision of 83%. This is the best model to predict recession. Based on this easily interpretable decision tree model following is observed:

A given quarter is predicted to be in Recession when:

- Last quarter year-over-year retail sales change is less than 38% (includes retail sales decreases as well) AND
- This quarter's monthly unemployment rate changes by more than 15% AND
- Last quarter drop in housing permits between high month and low month is greater than 39K AND
- Last quarter's year-over-year housing price index change is less than 11%.

References

[1] Data Source: Federal Reserve Economic Data [FRED] from St. Louis Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/

Acknowledgement

I am appreciative of the motivation, support and guidance provided by Dr. Badrinath Kottimukkular, Assistant Professor at George Washington University, Washington DC.