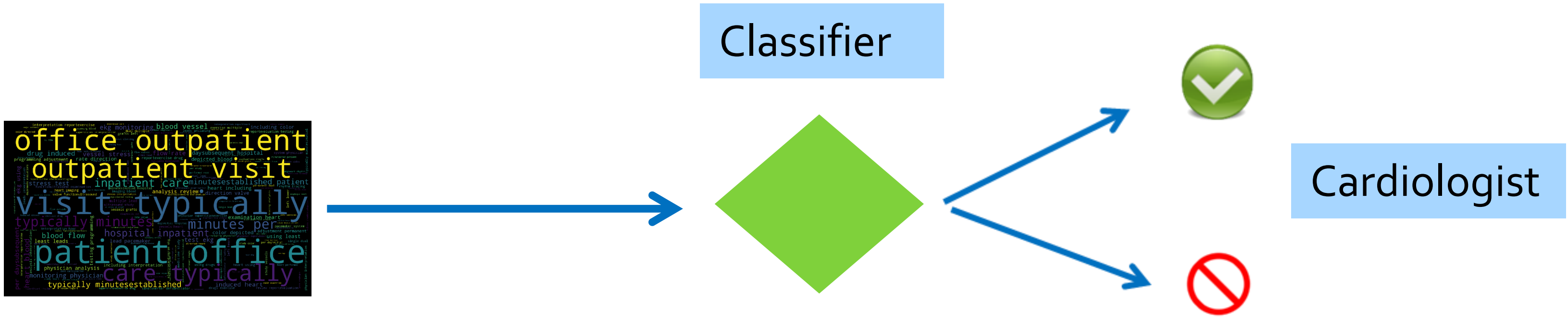


DOCREACH

PREDICTING PHYSICIAN SPECIALTY FROM TEXT DATA

(Data Confidential)
Sushil Sharma - Galvanize Cohort#65
Capstone Project – May'2018
linkedin.com/in/krishnatray
github.com/krishnatray

Business Problem - Summary

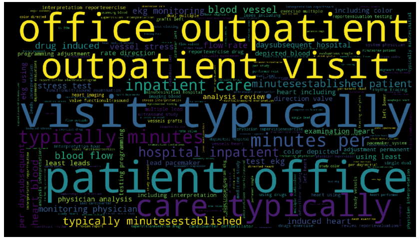
A social media marketing firm wants to **target doctors / physicians** based on their **practice area**.

Example: A **marketing campaign** to target **Cardiologists** for heart related news feed.

Business Problem

Sarah
Marketing Manager

Marketing Campaign
e.g. News Feed, Adv.



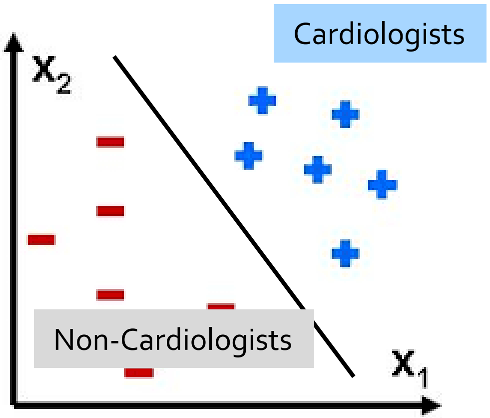
Cardiologists

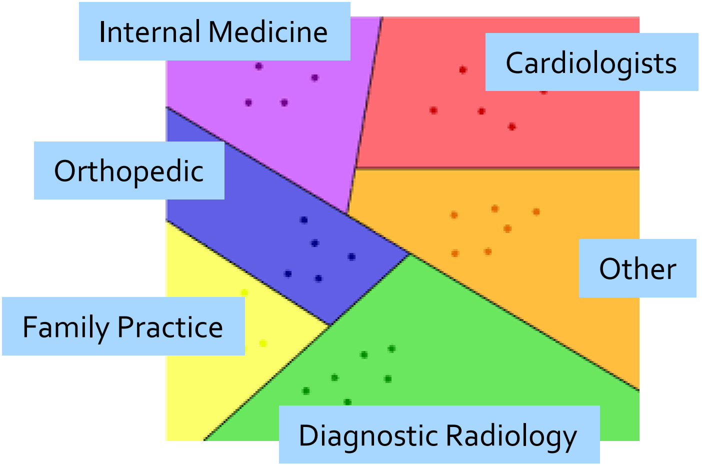
Example: A marketing campaign to target Cardiologists for heart related news feed.

Project Scope

1 Predict cardiologists vs Non-Cardiologists

2 Multiclass classifier to predict top 5 specialties





Data Source(s)

procedures.csv - contains a list of procedures doctors performed over the past year. The columns of this dataset are as follows:

- physician_id unique physician identifier, joins to id in physicians.csv
- procedure_code unique code representing a procedure
- **procedure** description of the procedure performed (Text Column)
- number_of_patients the number of patients the doctor performed that procedure on over the past year

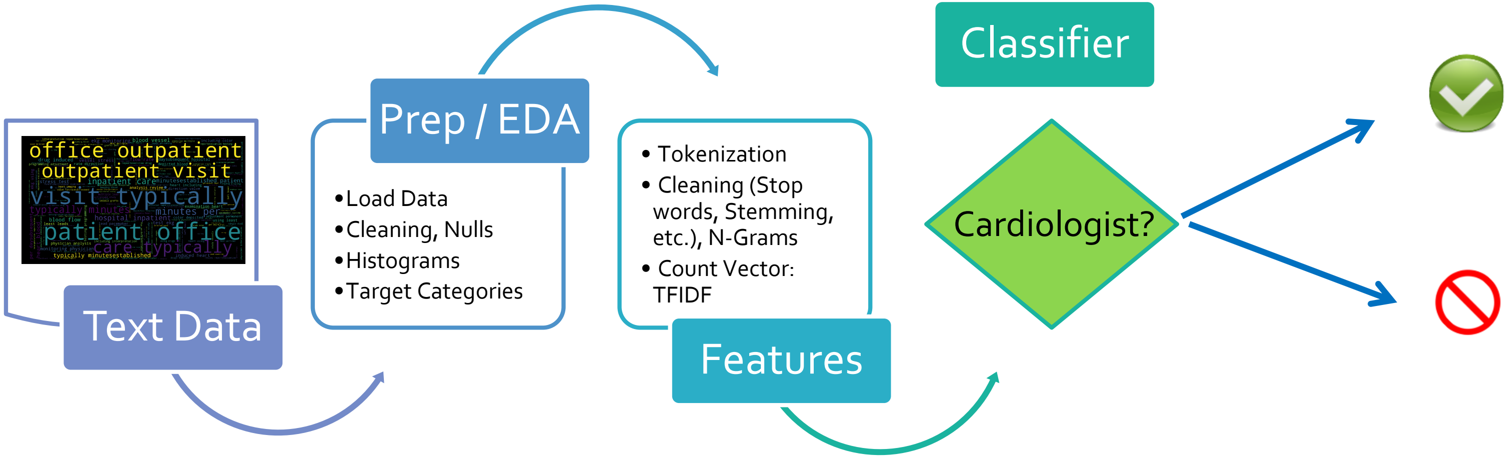
physicians.csv - contains a list of doctors and their unique specialty. Specialty is listed as "Unknown" for the doctors need to be classified.

- physician_id unique physician identifier, joins to id in physicians.csv
- **Specialty** String e.g. Cardiologist

Text Column

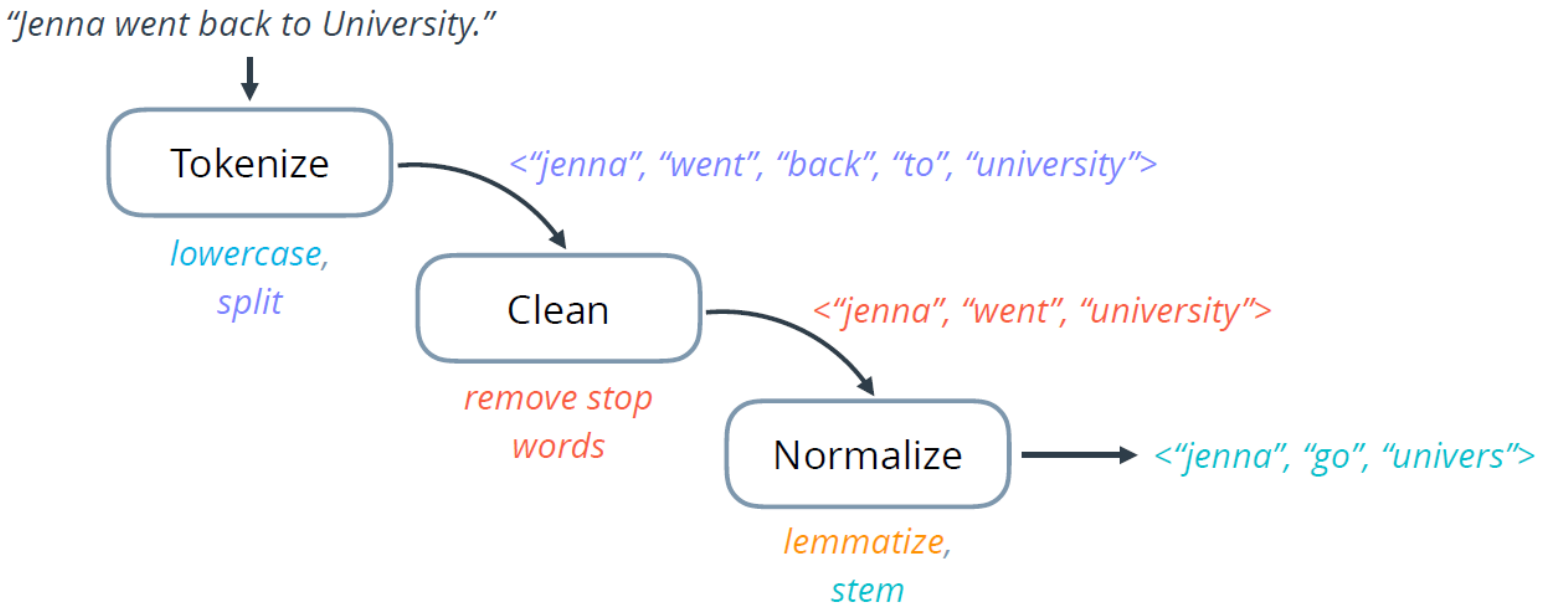
Category Column

Text Processing



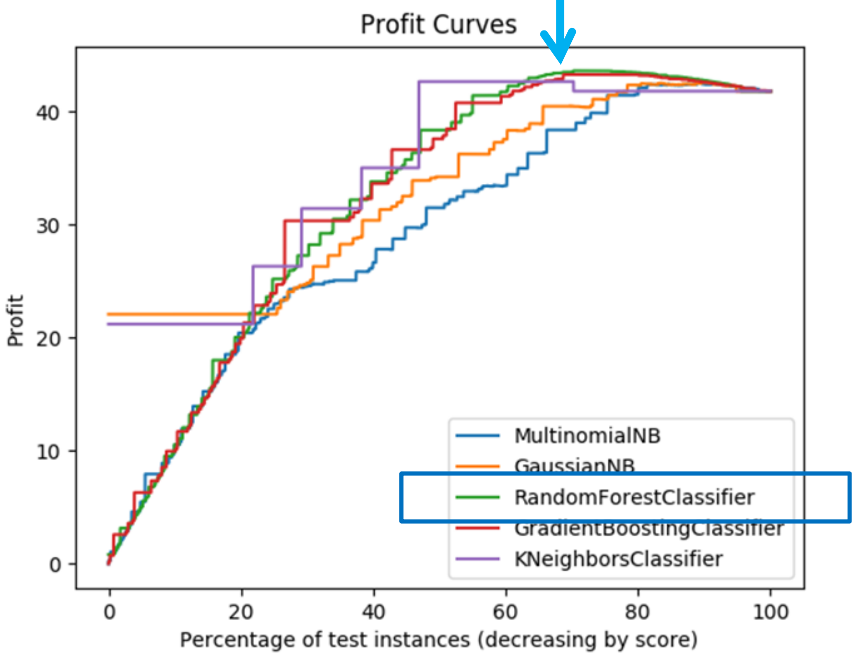
	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	0.18	**0.48**	0.18							
Doc 2	0.18		0.18	**0.48**	0.18	0.18				
Doc 3					0.18	0.18	**0.48**	**0.95**	**0.48**	**0.48**

Text Processing Summary



Results – Best Model (Binary)

- **Model Selection Method:** Highest Profit
- **Best model:** <u>Random Forest</u>
- **Resulting profit:** $43.64
- **Accuracy:** 80.5%
- **F1 Score :** 0.81

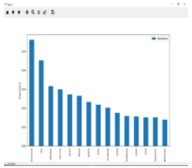


Results: Multiclass

Model Selection Method : Highest F1 Score

Model	Accuracy	Precision	Recall	F1-Score
MultinomialNB	66.5%	0.67	0.67	0.66
GrBoosting	69.8%	0.67	0.70	0.66
RandomForest	70.8%	0.68	0.71	0.68
XGBoost	68.8%	0.65	0.69	0.65

Top Features – Binary vs Multiclass



Top features – Binary Classifier	
1	**heart**
2	interpretation
3	study
4	blood
5	**ekg**

Top features – Multiclass Classifier	
1	**ultrasound**
2	ray
3	**minutes**
4	vaccine
5	heart

Next Steps / Future Plans

Unsupervised Machine Learning	Clustering, PCA, T-SNE
Deep Learning	Explore Deep Learning for Text Classification
Model Performance	Find additional Features
	Custom Ensemble Models
API / APP	Create Flask / Javascript Application /Api

Learnings / Challenges

High Dimensionality

- Training is Slow
- Limited features to 1000

No free lunch theorem

- RandomForest outperfomed other algorithms (Binary classification)

Gradient boosting is slow (Sequential)

- XGBoost

Data Volume

- Parameter to train on sample
- AWS machine with multiple CPU

SSH session timeouts / Process Monitoring

- Tmux
- htop

Technology Used

python™

NumPy

SciPy

matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn

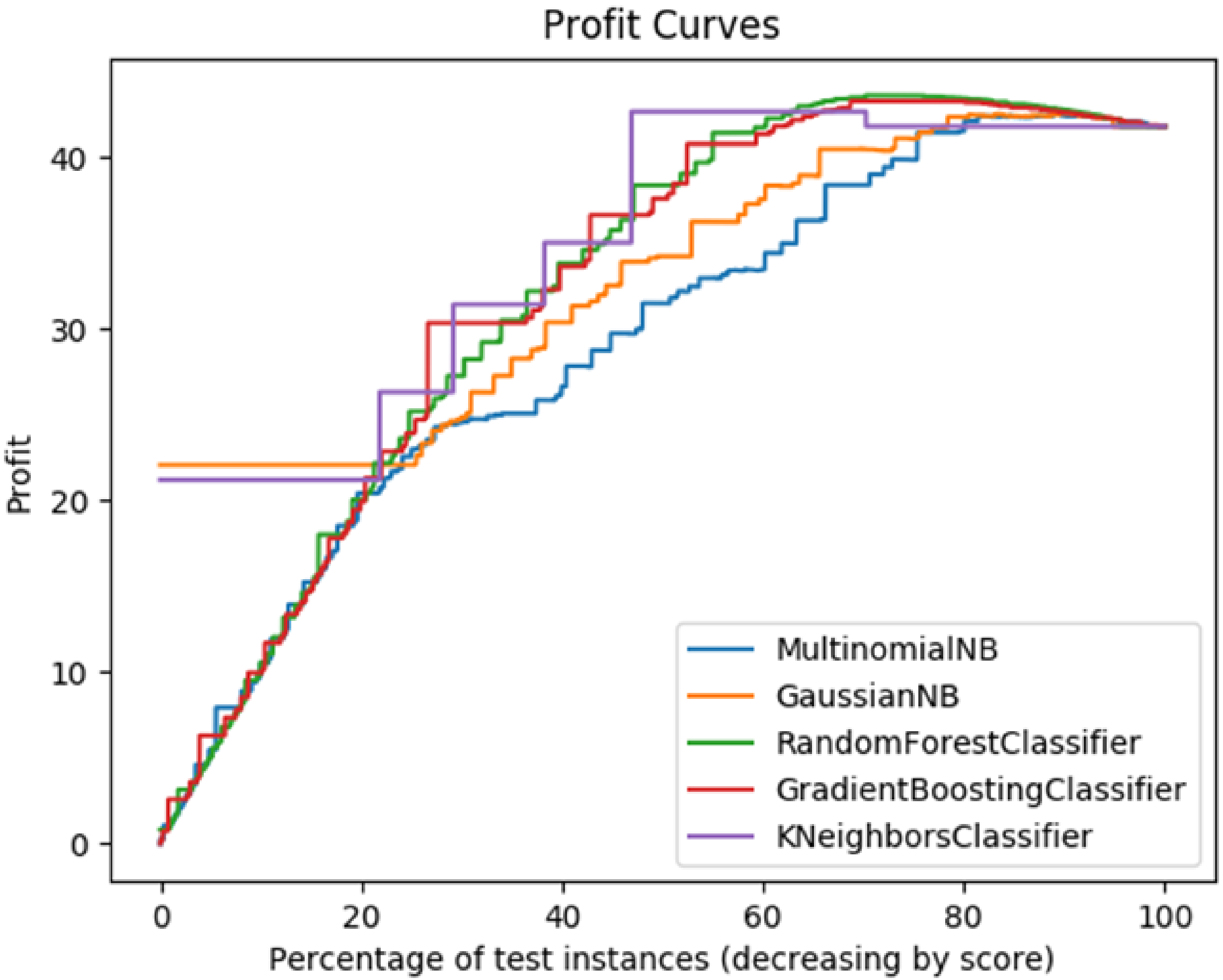
GitHub

amazon web services

THANKS !

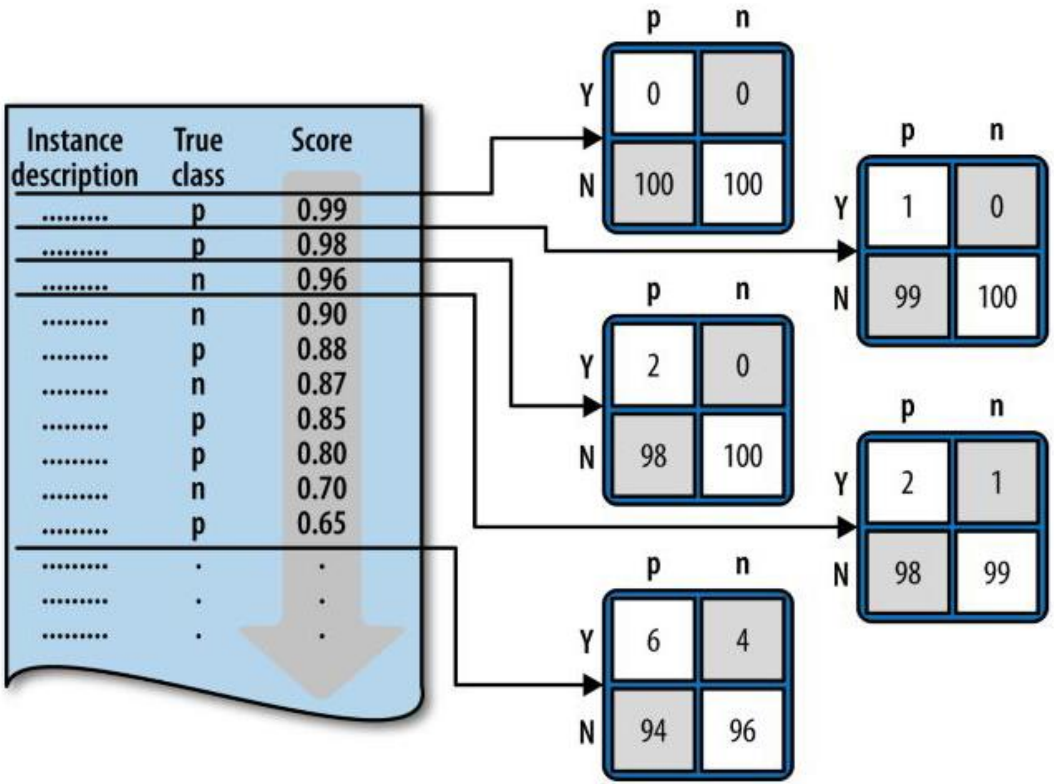
Sushil Sharma
Linkedin: https://www.linkedin.com/in/krishnatray
Github: https://www.github.com/krishnatray

BACKUP SLIDES


Profit Curves
Profit
Percentage of test instances (decreasing by score)
0
10
20
30
40
0
20
40
60
80
100
MultinomialNB
GaussianNB
RandomForestClassifier
GradientBoostingClassifier
KNeighborsClassifier

Cost Benefit

		Predicted	
		1	0
Actual	1	TP $100	FN = 0
	0	FP -$10	TN = 0



Data Summary

- Total Instances: 406,690
- Procedure length varies 6 – 256; Median 66
- Binary Classes
  - Cardiologists 47%
  - Non-Cardiologist 53%
- Multiclass picked top 5 and combined rest specialty as other
  - Cardiology 47%
  - Diagnostic Radiology 10%
  - Family Practice 7%
  - Internal Medicine 9%
  - Orthopedic Surgery 3%
  - Other 26%

Term Frequency

Normalize counts within a document to frequency

$$tf(t,d) = \frac{total\ count\ of term\ t\ in\ document\ d}{total\ count\ of\ all\ terms\ in\ document\ d}$$

document	galvanize	learn	other	student	teach
Doc 1	0	¼ = 0.25	¼ = 0.25	2/4 = 0.5	0
Doc 2	½ = 0.5	0	0	0	½ = 0.5
Doc 3	⅓ = 0.33	⅓ = 0.33	0	⅓ = 0.33	0

Inverse Document Frequency

$$idf(t, D) = log\frac{total\ number\ of\ document\ incorpus\ D}{count\ of\ document\ containing\ term\ t}$$

document	galvanize	learn	other	student	teach
Doc 1		X	X	X	
Doc 2	X				X
Doc 3	X	X		X	
idf(t,D)	log(3/2)	log(3/2)	log(3/1)	log(3/2)	log(3/1)

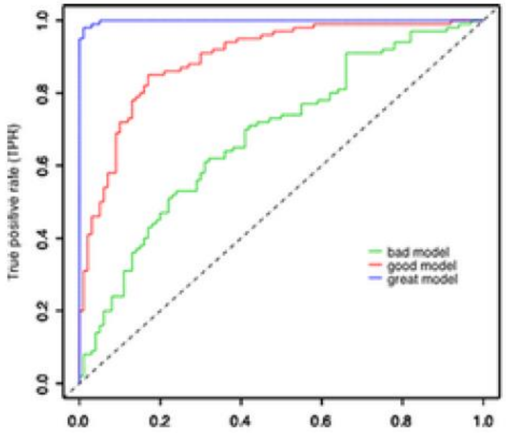
TF-IDF

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

document	galvanize	learn	other	student	teach
Doc 1	0	0.25xlog(3/2) = 0.101	0.25xlog(3/1) = 0.275	0.5xlog(3/2) = 0.203	0
Doc 2	0.5xlog(3/2) = 0.203	0	0	0	0.5xlog(3/1) = .549
Doc 3	0.33xlog(3/2) = 0.135	0.33xlog(3/2) = 0.135	0	0.33xlog(3/2) = 0.135	0

ROC Curve

- A plot of the TPR vs. FPR at difference thresholds is called a ROC curve. It is used to visualize the performance of a given binary classifier:



- **Accuracy** - How many observations did I label correctly?

$$\frac{TP + TN}{P + N}$$

- **True Positive Rate (TPR), Recall, Sensitivity** - Of those observations that are actually positives, which ones did I label as positive?

$$\frac{TP}{TP + FN}$$

- **False Positive Rate (FPR)** - Of those observations that are actually negatives, which ones did I label as positive?

$$\frac{FP}{FP + TN}$$

Classification Metrics

- **Precision, Positive Predictive Value** - Of those observations that I labeled as positive, which ones are actually positive?

$$\frac{TP}{TP + FP}$$

- **True Negative Rate, Specificity** - Of those observations that are actually negative, which ones did I label as negative?

$$\frac{TN}{TN + FP}$$

	Predicted 1	0
Actual 1	TP	FN
0	FP	TN

Precision = P(Actual 1 | Predicted 1) = TP / (TP + FP)

Recall Or TPR = P(Prediction 1 | Actual 1) = TP / (TP + FN)

EDA

	physician_id	procedure_code	procedure	number_of_patients
0	0	99202	new_patient_office_or_other_outpatient_visit_...	14
1	0	99203	new_patient_office_or_other_outpatient_visit_...	15
2	0	99205	new_patient_office_or_other_outpatient_visit_...	12
3	0	99212	established_patient_office_or_other_outpatient...	27
4	0	99213	established_patient_office_or_other_outpatient...	16

	id	specialty
0	0	General Surgery
1	1	Unknown
2	2	Family Practice
3	3	Emergency Medicine
4	4	Plastic and Reconstructive Surgery

```
In [30]: top5 = data['target'].value_counts()[:5]
         top5

Out[30]: Cardiology               190095
         Diagnostic Radiology      39217
         Internal Medicine         36357
         Family Practice           26729
         Orthopedic Surgery        10276
         Name: target, dtype: int64
```

Common words