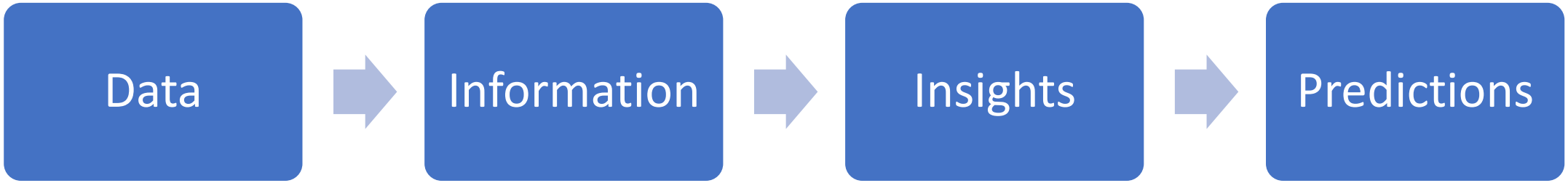
A Technique for Improving Classification Accuracy of Highly Imbalanced and Sparse Datasets

Sindhura Bonthu

Advisor – Dr.Qiuming Zhu

Big Data and Machine Learning

- Big Data – Massive volumes of data that is often characterized with 3 Vs – volume, variety, and velocity

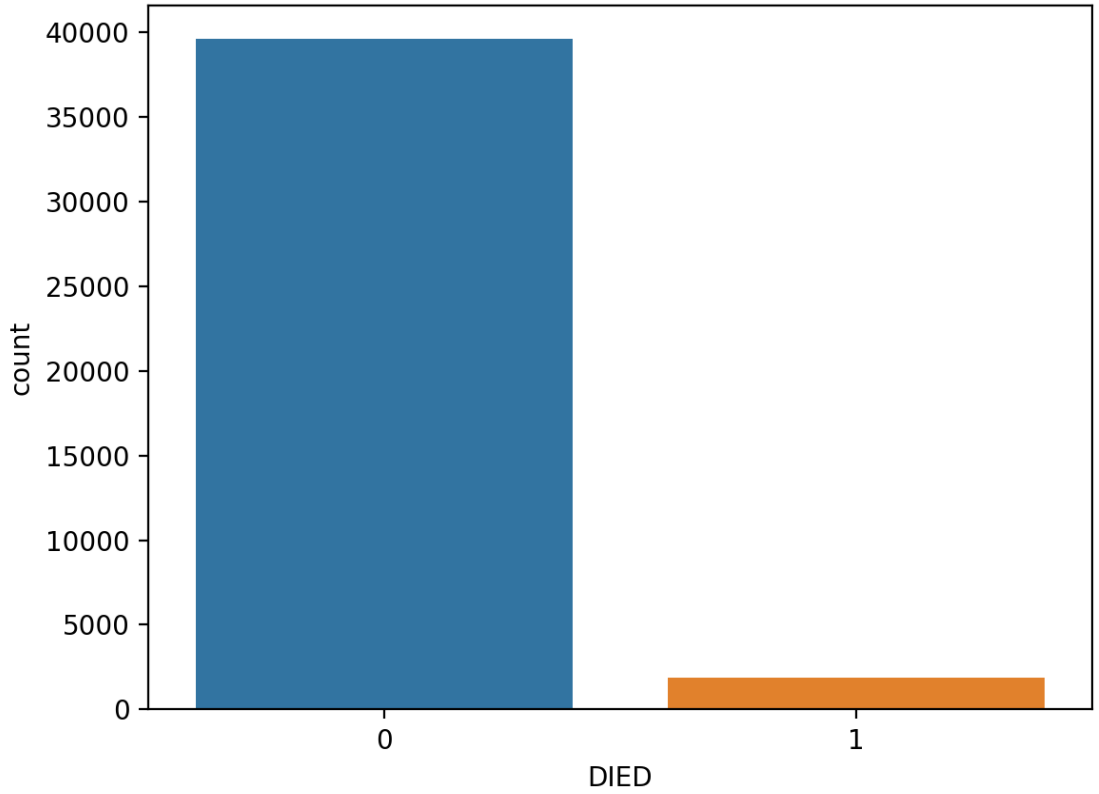


- Machine Learning – Field of study that gives computers the ability to learn without being explicitly programmed

Challenges

- Data Sparsity – Data is considered sparse when certain expected values in a dataset are missing or not present
- Data imbalance – Data is said to be imbalanced when we have multiple classes of data but the majority of the data belongs to one class
- Examples
  - Fraud detection in banking application
  - Medical diagnosis of a rare disease
  - Detection of oil spill in satellite radar image

Data Imbalance



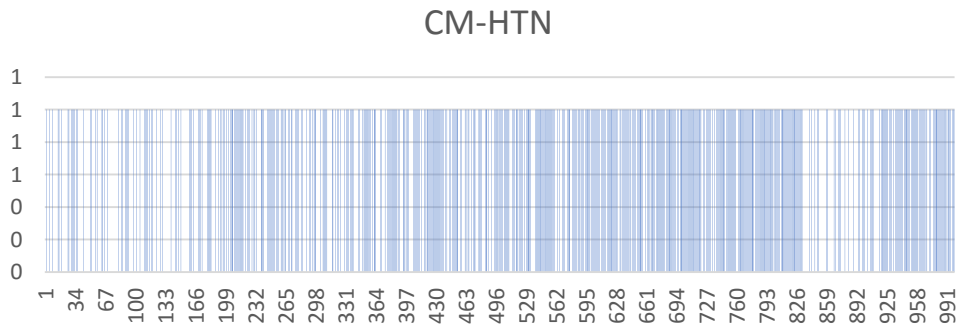
Classification of Colorectal Data based on the column ‘Died’

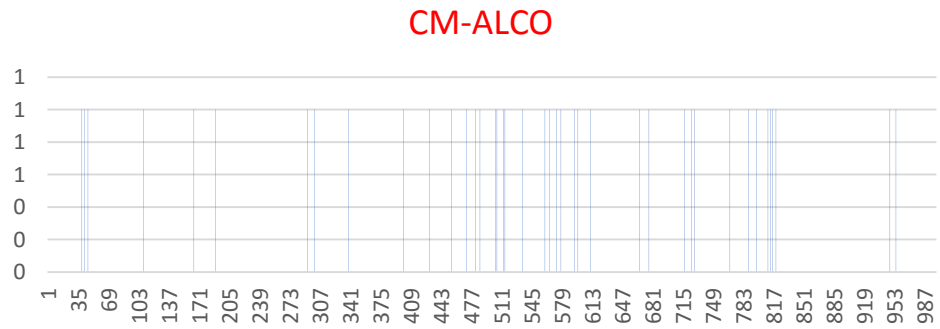
Dealing with imbalanced Dataset

- Data Preprocessing
- Using the right evaluation metric
- Data resampling strategies
- Algorithms
- Neural Network Architecture

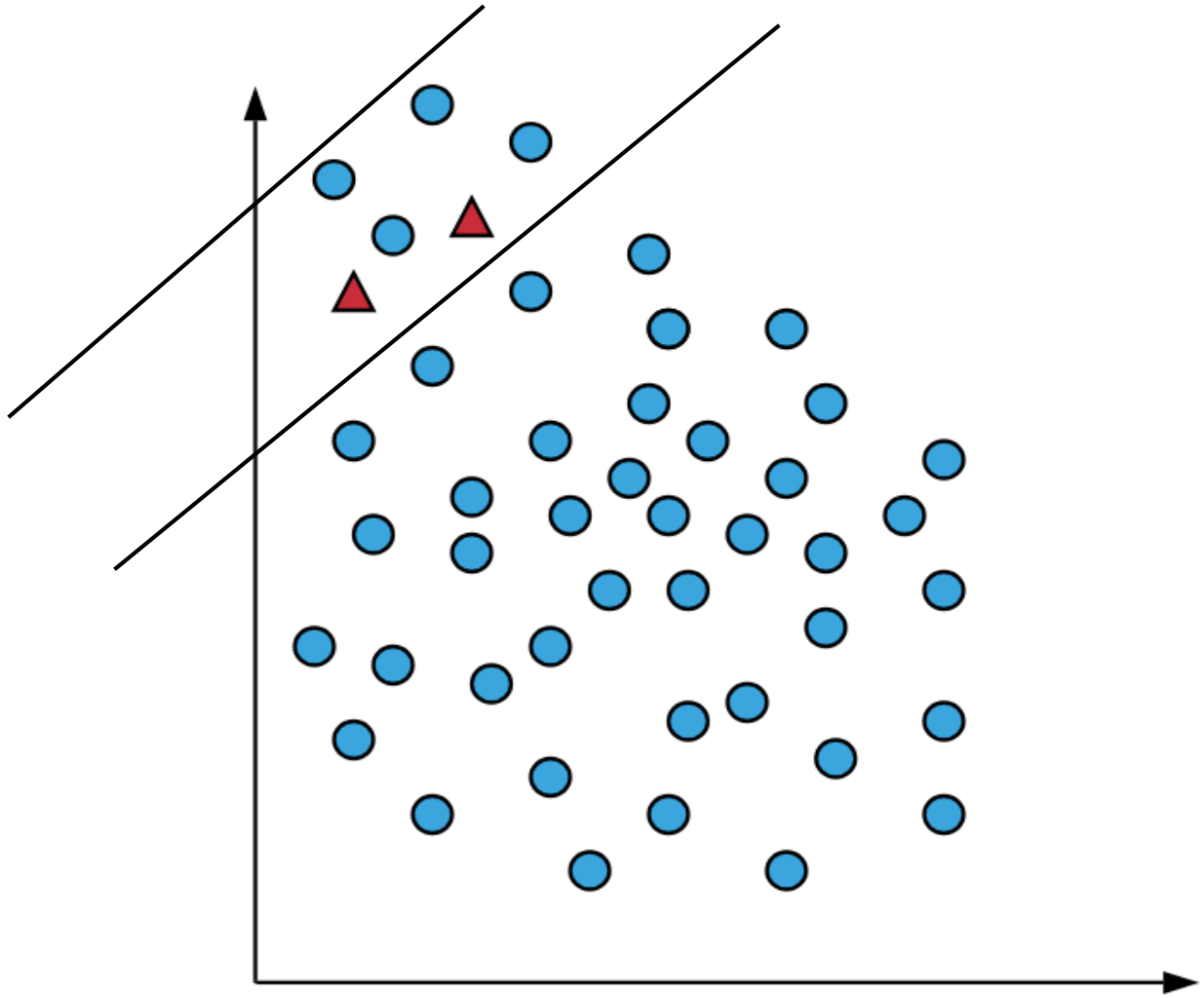
Data Pre-Processing

- Data Cleaning – Created a data file that removed all the cases with no no-zero valued fields
- Identifying most significant features – Statistic analysis on individual features





The Right Evaluation Metric



Confusion Matrix

- For a binary classification problem, confusion matrix is a 2X2 matrix which is defined as follows:

TN	FP
FN	TP

- Where TN – True Negatives, FP – False Positives, FN – False Negatives, TP – True Positives
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)

Data Resampling techniques

- Under Sampling – Under Sample the majority class

  Disadvantage – Misses important information
- Over Sampling – Over Sample the minority class

  Disadvantage – Over fitting problem
- Synthetic Minority Over Sampling – Over Sample by creating Synthetic samples of minority class

Algorithms

- Logistic regression with resampling techniques, Relu
- Ensemble composite models - combines series of low performing classifiers to create improved classifier
- Boosting algorithms

  Builds a model from the training data

  Creates a second model that attempts to correct the errors from the first model

  Models are added until the training set is predicted perfectly

Neural Network Architectures

- Experiment with different neural network architectures by varying number of layers, hidden nodes to find the optimal result.

Model	# layers	#Nodes for each layer				# Training samples	# Validation samples	Training Data		Validation Data	
		L1	L2	L3	L4			Loss	Accuracy	Loss	Accuracy
1	3	7	4	1		219	109	0.2003	88.55	0.2101	73.39
2	3	7	4	1		262	66	0.1061	88.55	0.0595	95.45
3	3	7	4	1		196	132	0.1545	87.24	0.086	93.18
4	3	8	4	1		219	109	0.1433	88.58	0.1093	88.07
5	3	8	4	1		262	66	0.109	88.55	0.0674	95.45
6	3	8	4	1		196	132	0.1466	87.24	0.0676	93.18
7	4	10	6	3	1	262	66	0.1067	88.55	0.0313	96.97
8	4	9	6	3	1	219	109	0.1049	88.58	0.0676	93.58
9	4	7	5	3	1	219	109	0.1173	88.58	0.0785	93.58
10	4	7	5	3	1	262	66	0.1016	88.55	0.0325	96.97

Next Steps

- Combination of techniques – combine all the above techniques in different experimental settings to improve the precision and recall of the classification data

THANK YOU
ANY QUESTIONS