Machine Learning for Imbalanced Class Distributions.

Tanisha R. Bhayani
(Associate AI Researcher @ F(x) Data Labs Pvt. Ltd.)

GDG Ahmedabad

Women Techmakers

“There’s nothing artificial about AI...It’s inspired by people, it’s created by people, and—most importantly—it impacts people. It is a powerful tool we are only just beginning to understand, and that is a profound responsibility.”

- Fei-Fei Li

(Chief Scientist of AI/ML of Google Cloud,

Professor Director, Stanford AI Lab Computer Science Department)

What is AI?

- Algorithmically, AI is the about solving those problems which are NP-hard.
- Time and Space Tradeoff.
- Human and AI. (Philosophically and Professionally)
- Do AI fail? - correctness of AI.
- IA - Intelligence Augmentation.

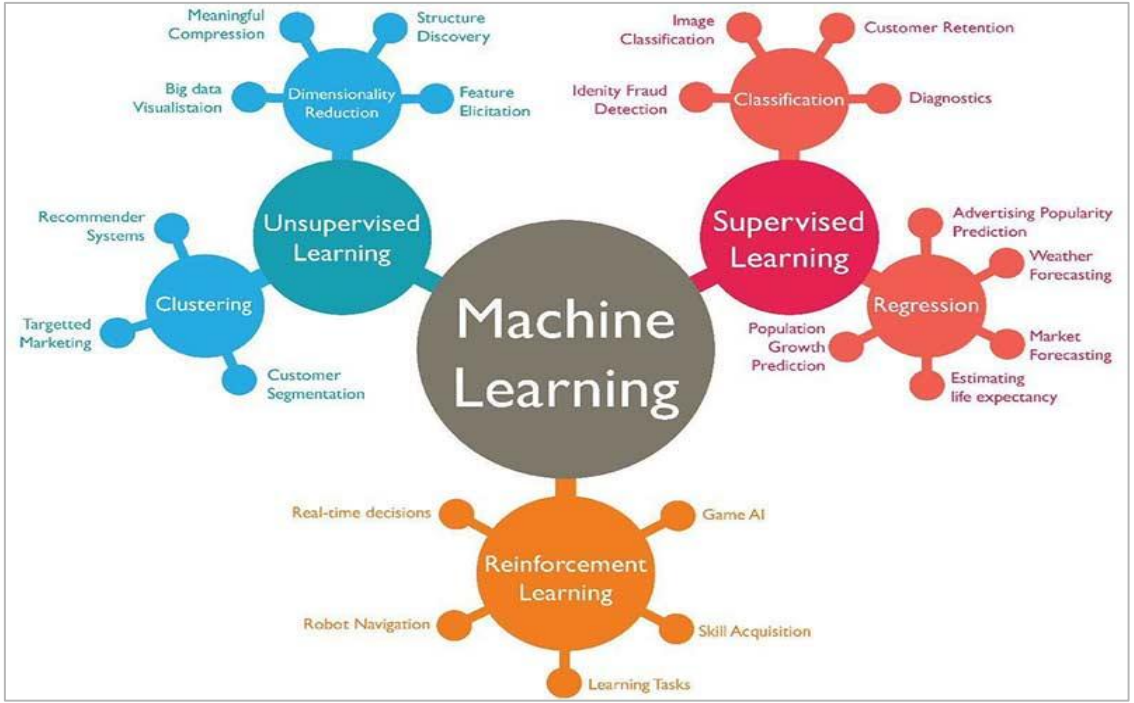
What is Machine Learning

- Why Machine Learning?
- What makes machine learning so powerful?
- Is everything just dependant on Machine Learning.

DEEP LEARNING

- Life is deep, so are neural networks.
- The way brain neuron learns.
- Inspiration.
- Old School AI.

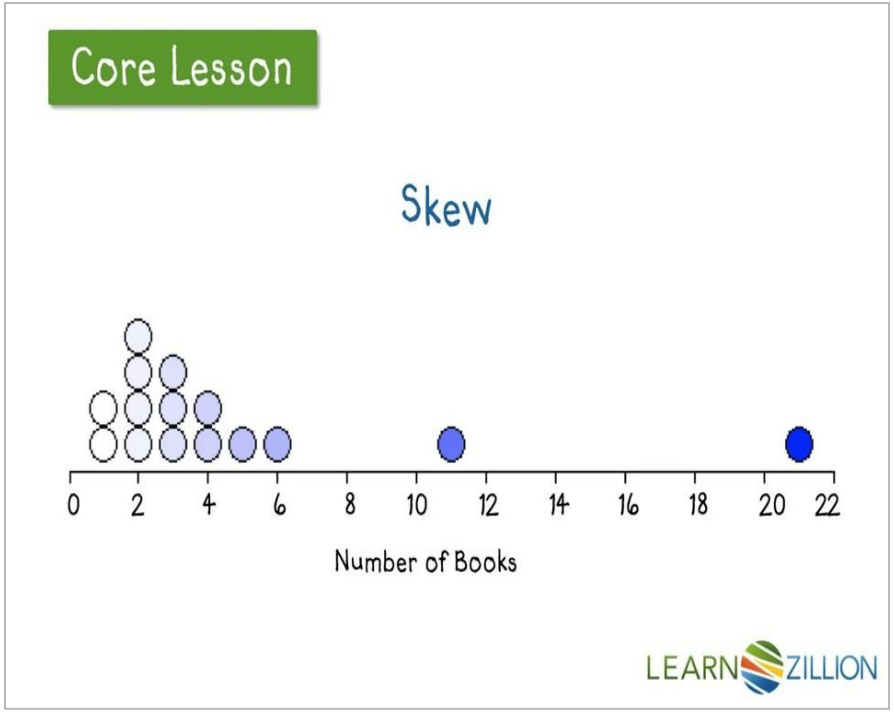
Types of Machine Learning



- Works on properties of data.
- The interaction of data with environment.
- The way the algorithm is designed.

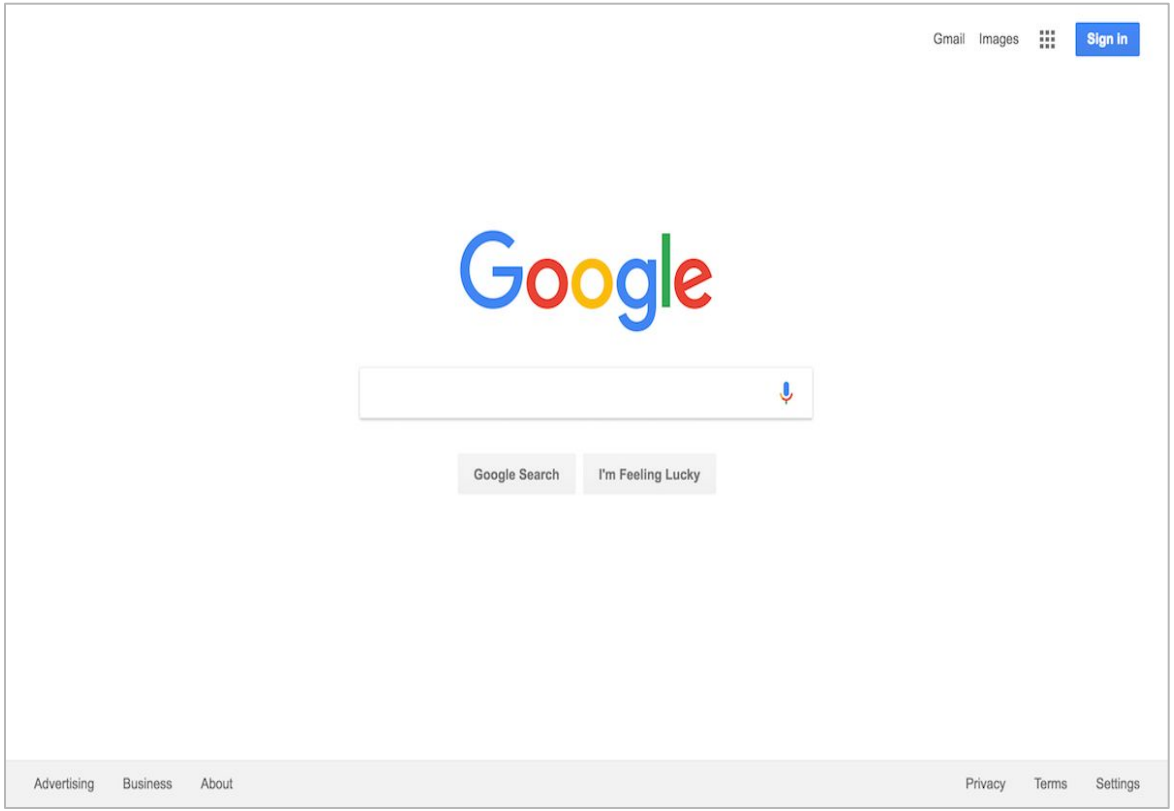
Kind of data required for Classification

- Labelled data (Long shot process)
- Balanced data
- Clean data
- Data having all the information
- Proper data distribution
- Different types of Learning for doing Classification



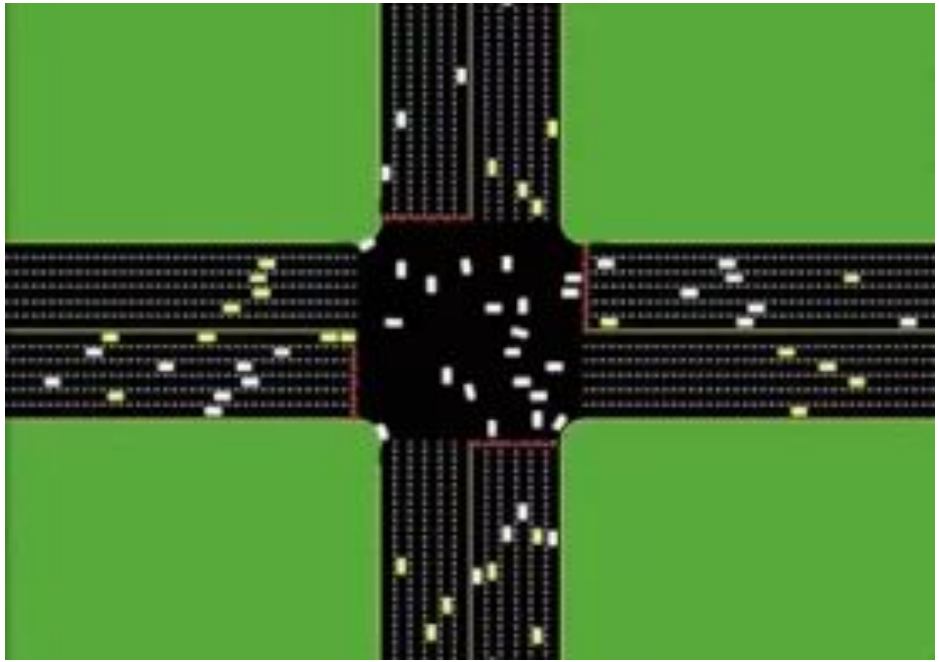
Are all the data in real world balanced?

Google Search Engine



- Query Matching
- Symbolic AI
- Deep Learning
- Page Rank

Self Driving Cars



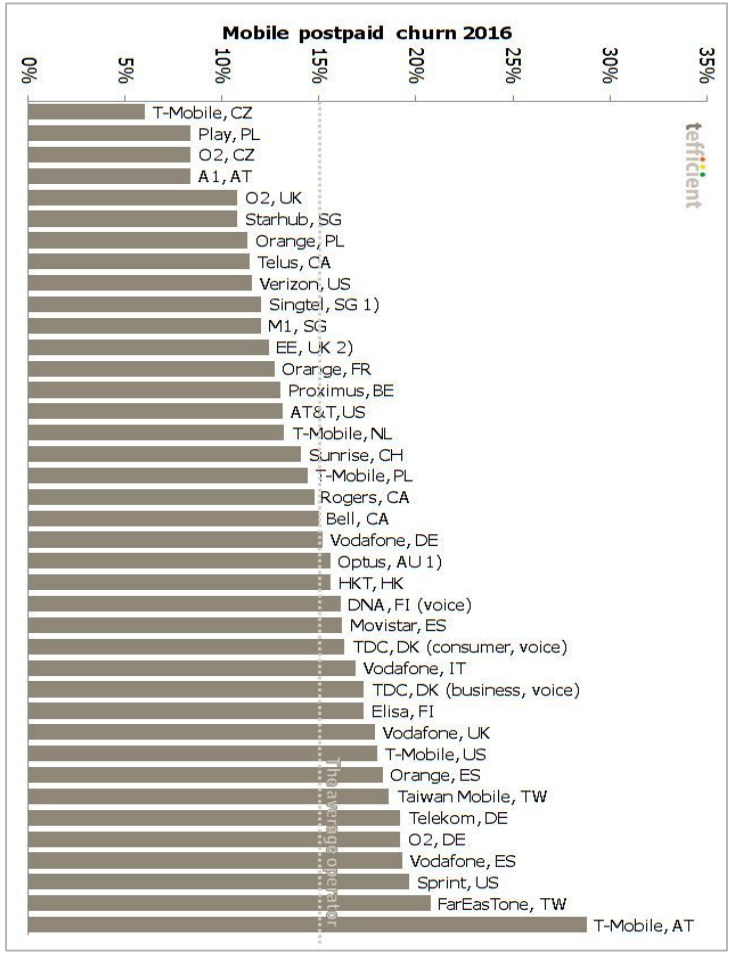
- Nash Equilibrium
- What should be considered as an obstacle
- Car as an entity
- Rare conditions which might occur

Part of Speech Tagging

```
[('Each', 'DT'),
 ('of', 'IN'),
 ('us', 'PRP'),
 ('is', 'VBZ'),
 ('full', 'JJ'),
 ('of', 'IN'),
 ('stuff', 'NN'),
 ('in', 'IN'),
 ('our', 'PRP$'),
 ('own', 'JJ'),
 ('special', 'JJ'),
 ('way', 'NN')]
```

Credit Card Fraud

Marketing



More Sophisticated Conversion Funnel

100%	Website visitors
10%	Sign-up to trial
9%	Login to account
7%	Use the key features
5%	Active users
1%	Paying users
0.8%	Staying users (>12 months)

Cuisine From Ingredients

```
df.cuisine.value_counts()
```

```
italian          7838
mexican          6438
southern_us      4320
indian           3003
chinese          2673
french           2646
cajun_creole     1546
thai             1539
japanese         1423
greek            1175
spanish           989
korean            830
vietnamese        825
moroccan          821
british           804
filipino          755
irish             667
jamaican          526
russian           489
brazilian         467
Name: cuisine, dtype: int64
```

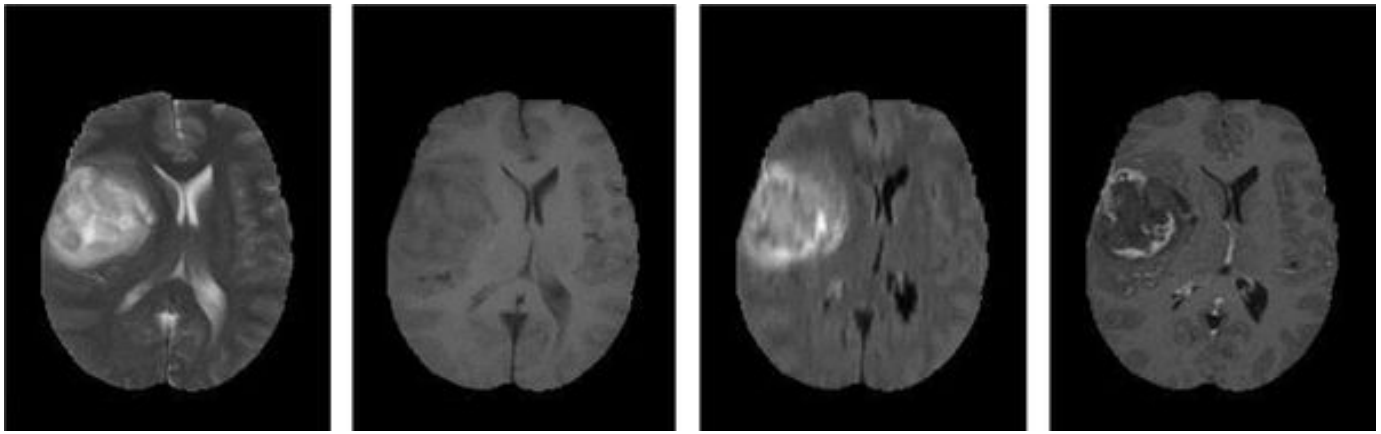
```
new_df['cuisine_%']
```

```
italian          19.706341
mexican          16.186453
southern_us      10.861367
indian            7.550158
chinese           6.720471
french            6.652587
cajun_creole      3.886961
thai              3.869362
japanese          3.577714
greek             2.954191
spanish           2.486549
korean            2.086790
vietnamese        2.074219
moroccan          2.064163
british           2.021421
filipino          1.898225
irish             1.676975
jamaican          1.322472
russian           1.229446
brazilian         1.174134
Name: cuisine_%, dtype: float64
```

https://www.kaggle.com/c/whats-cooking/data

Medical Diagnosis

Brain Tumour Identification



Bias and Prejudice

- GIGO
- Data collection practices
- Only patterns are collected and not user information
- Computer generated or human created?
- Decisions based on features.
- Not all features are covered.

Algorithms and data sampling methods required for handling skew data.

- Importance of data or algorithms
- Correctness of both
- Time analysis

Data Sampling

1. Under Sampling
2. Over Sampling
3. Creating Synthetic data - SMOTE (Synthetic Minority Over-Sampling Technique)

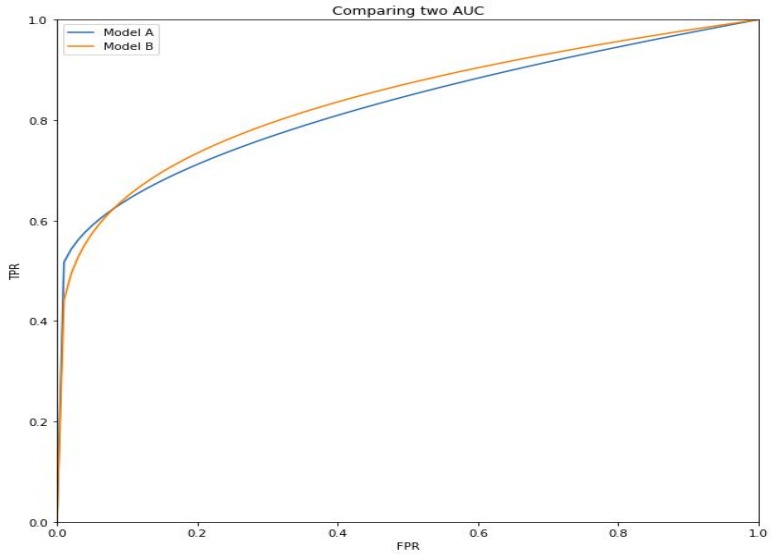
Algorithms

1. Cost Sensitive Learning
2. Modified SVM
3. KNN
4. Neural Networks
5. Genetic Programming
6. Probabilistic Decision Tree
7. Rough Set based methods
8. Bagging
9. Boosting

Testing These Models

1. Accuracy
2. True Positive Rate, False Positive Rate - AUROC
3. Geometric Mean Score
4. Confusion Matrix
5. Threshold Decision

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN



Current Research Trends in handling skew data.

1. Reinforcement Learning Algorithms
2. Algorithms for Multiclass Classification
3. Deep Learning

Implementation of various methods

1. Sampling methods
2. Cost sensitive Learning
3. Conventional Machine Learning model on dataset

Feature Engineering

- What is feature engineering?
- What do we recognize?
- Should all features be in same reference system?
- Data normalization
- Why is it important?

Creating Synthetic Features

Creating new information from existing information

How to do that?

Domain Knowledge?

Human Inference knowledge.

Implementation and Questions

CODING

https://github.com/RoshanTanisha/ML_GDG

QUESTIONS

FEEDBACK

THANK YOU!