Machine Learning using Python

Manaranjan Pradhan

About Me

- Consulting and training on Big data, AI & Machine Learning.
- An alumni of *IIM, Bangalore*.
- Has about 20+ years of industry experience.
- Has trained 1000+ professionals on Big Data and AI & ML.
- An adjunct faculty at IIM, Bangalore, ISB, Hyderabad and Jio Institute

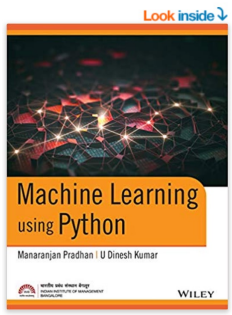
Manaranjan Pradhan | LinkedIn

LinkedIn: https://in.linkedin.com/in/manaranjanpradhan
Personal Website: https://www.manaranjanp.com/

Manaranjan has co-authored the best-selling book Machine Learning using Python

He has published the following machine learning cases in **(HBR) Harvard Business Publishing**:

1. Customer Analytics at Big Basket – Product Recommendations
2. Improving Lead Generation at Eureka Forbes Using Machine Learning Algorithms



Machine Learning using Python Paperback – 2019

by U Dinesh Kumar Manaranjan Pradhan (Author)

7 customer reviews

See all 2 formats and editions

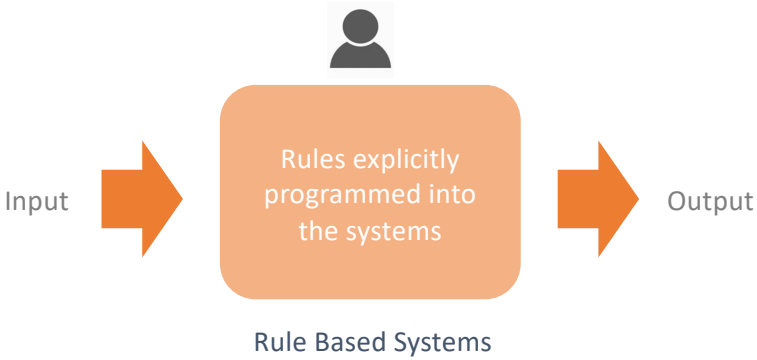
Kindle Edition ₹ 423.20 | Paperback ₹ 529.00 prime

Read with Our Free App | 2 New from ₹ 529.00

This book is written to provide a strong foundation in machine learning using Python libraries by providing real-life case studies and examples. It covers topics such as foundations of machine learning, introduction to Python, descriptive analytics and predictive analytics. Advanced machine learning concepts such as decision tree learning, random forest, boosting, recommended systems, and text analytics are covered. The book takes a balanced approach between theoretical understanding and practical applications. All the topics include real-world examples and provide step-by-step approach on how to explore, build, evaluate, and optimize machine learning models.

https://www.amazon.in/Machine-Learning-Python-Manaranjan-Pradhan-ebook/dp/B07RLQPNRX

Rule Based or Expert Systems



Static Rules:

- If a user's credit card country points to the US but their IP points to Russia, then the transaction should be blocked.

Velocity Rules:

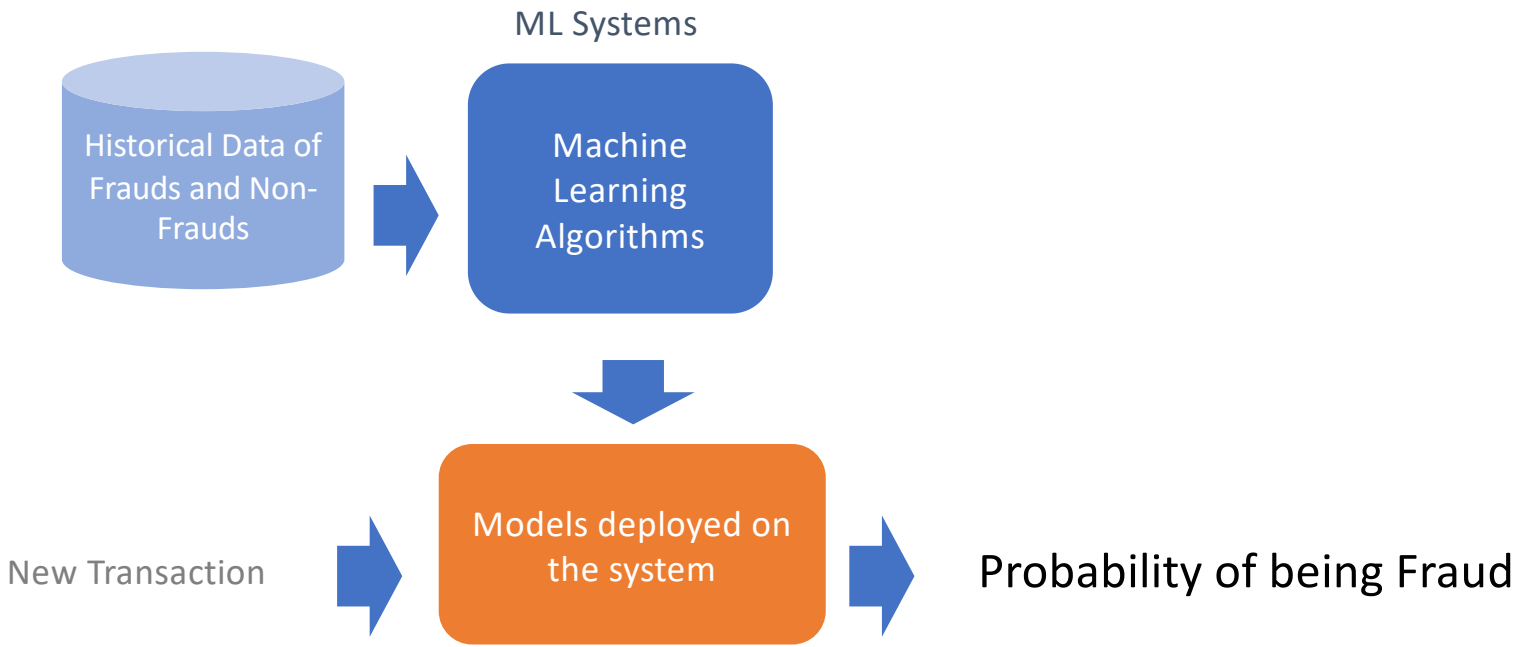
These rules attempt to **understand user behaviour by looking at set actions over a time period**.

- An increase in spending (more than 200%) over a 24-hour period
- A single user attempting to pay with five different frozen credit cards *within ten minutes* is highly suspicious, as even someone in dire straits would likely stop once they realize one or two of their cards have been frozen or cancelled.

https://seon.io/resources/guides/guide-to-fraud-detection-rules/

Limitation of Rule Bases Systems

Machine Learning Systems

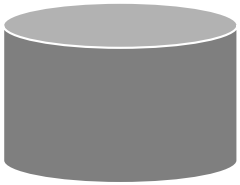


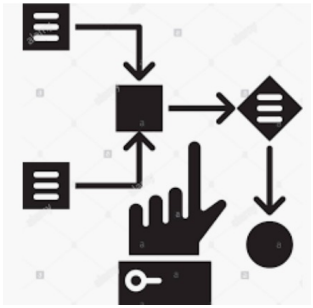
What is Machine Learning?

Machine learning is a field of study that gives computers the ability to learn without explicitly being programmed.

Source: MIT Sloan

Key elements of Machine Learning





$$y = f(x1, x2, x3....)$$

Data

- Past experiences
- Samples representing problem context

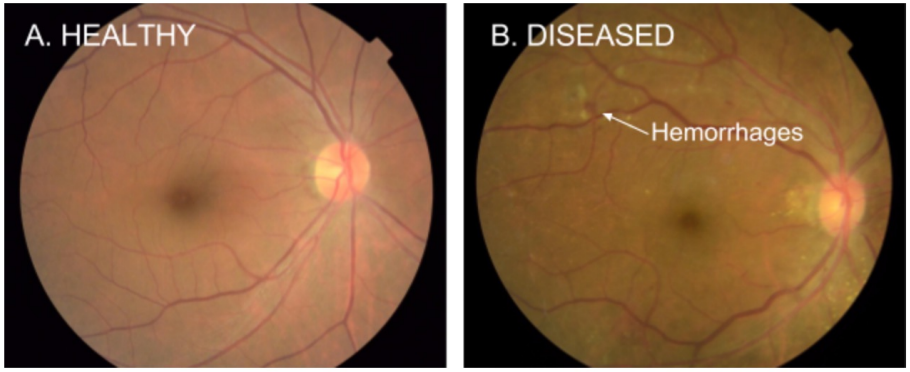
Algorithms

- Machine Learning
- Iteratively goes through the data to find a pattern in the data

Model

- A mathematical expression or set of rules representing the the pattern found in the data

Detection of Diabetic Eye Disease



haemorrhage
/ˈhɛmərɪdʒ/

noun
plural noun: **hemorrhages**

1. an escape of blood from a ruptured blood vessel.
   "a massive haemorrhage of the brain"

https://research.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html

- Google working closely with doctors both in India and the US, created a development dataset of 128,000 images which were each evaluated by 3-7 ophthalmologists from a panel of 54 ophthalmologists.
- Trained a deep neural network to detect referable diabetic retinopathy.
- Then tested the algorithm's performance on two separate clinical validation sets totaling ~12,000 images, with the majority decision of a panel 7 or 8 U.S. board-certified ophthalmologists serving as the reference standard.
- the algorithm has a F-score (combined sensitivity and specificity metric, with max=1) of 0.95, which is slightly better than the median F-score of the 8 ophthalmologists we consulted (measured at 0.91).

Customers who bought this item also bought…

Into Thin Air: A Personal Account of the Mt. Everest Disaster

Paperback – October 19, 1999

by Jon Krakauer (Author, Photographer), Randy Rackliff (Illustrator), Daniel Rembert (Contributor), & 2 more

2,414 customer reviews

#1 Best Seller in Mountain Climbing

See all 65 formats and editions

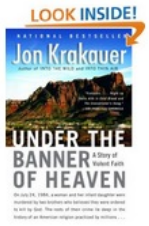
Kindle	Hardcover	Paperback	Mass Market Paperback
$3.40	$18.66	$10.36	from $0.01
Read with our free app	667 Used from $0.01 82 New from $4.49 51 Collectible from $6.37	483 Used from $0.01 117 New from $6.22 11 Collectible from $9.70	487 Used from $0.01 18 New from $4.59 19 Collectible from $3.00

National Bestseller

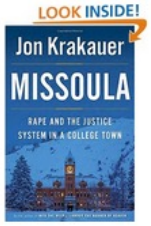
A bank of clouds was assembling on the not-so-distant horizon, but journalist-mountaineer Jon

Customers Who Bought This Item Also Bought

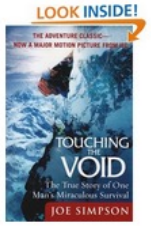
Into the Wild
› Jon Krakauer
2,304
#1 Best Seller in Travelogues & Travel Essays
Paperback
$7.34 Prime



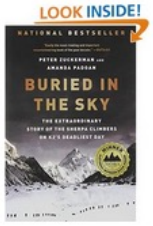
Under the Banner of…
› Jon Krakauer
1,361
Paperback
$10.03 Prime
Get it by Tomorrow



Missoula: Rape and the…
Jon Krakauer
361
Hardcover
$18.09 Prime
Get it by Tomorrow



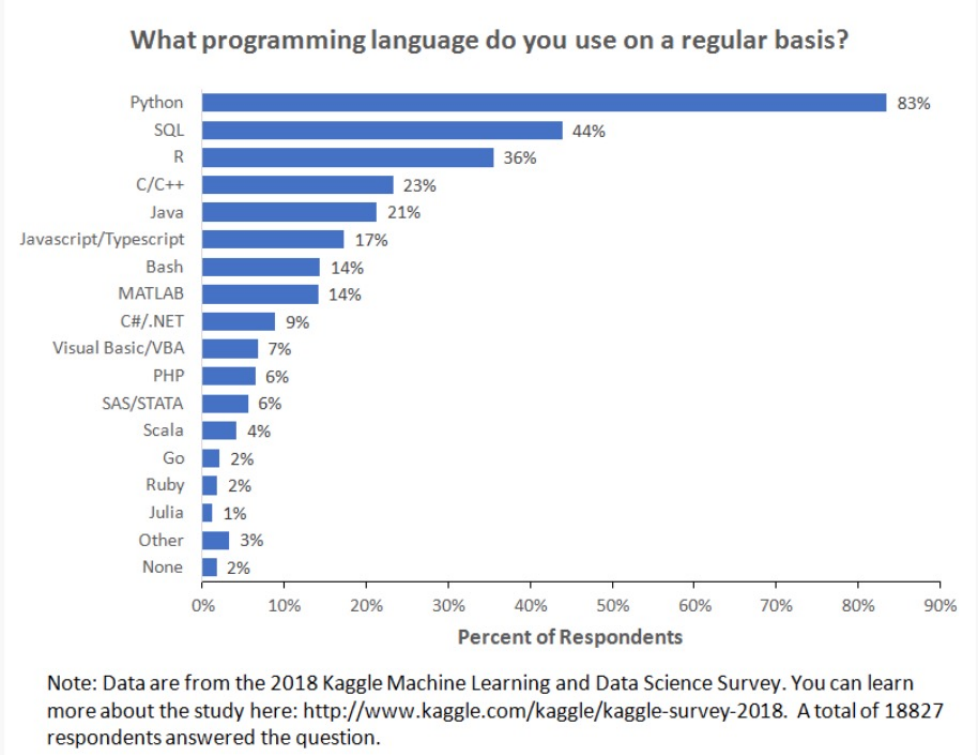
Touching the Void: The…
› Joe Simpson
315
Paperback
$11.22 Prime
Get it by Tomorrow



Buried in the Sky: The…
› Peter Zuckerman
225
Paperback
$10.63 Prime
Get it by Tomorrow

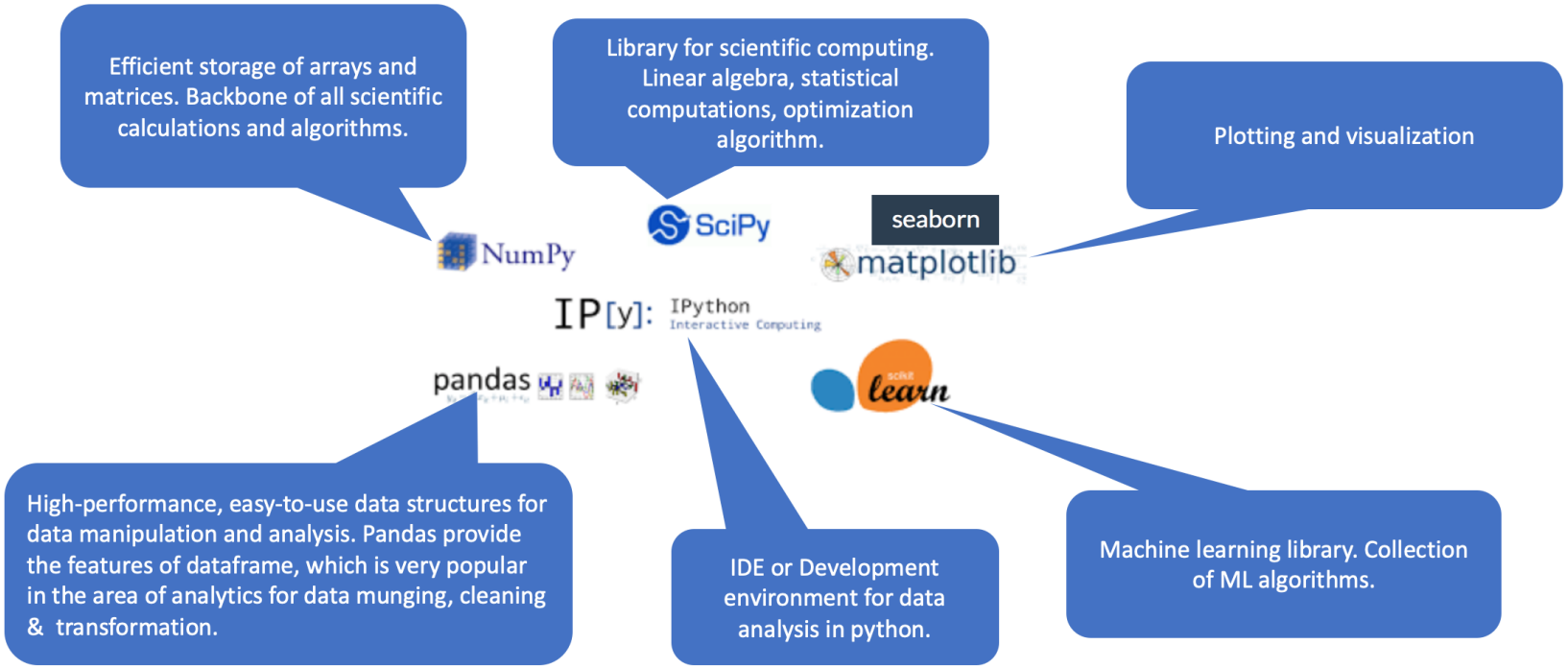
What are key elements of Machine Learning System?

Language for Machine Learning



https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/

Python Stack For Data Science



Development Environment and Tools

The Jupyter Notebook is **a web-based interactive computing platform**.

https://www.jupyter.org/

Github

The Jupyter Notebook is **a web-based interactive computing platform**.

https://www.github.com/

Platforms

ANACONDA.

Most popular open-source Python distribution platform

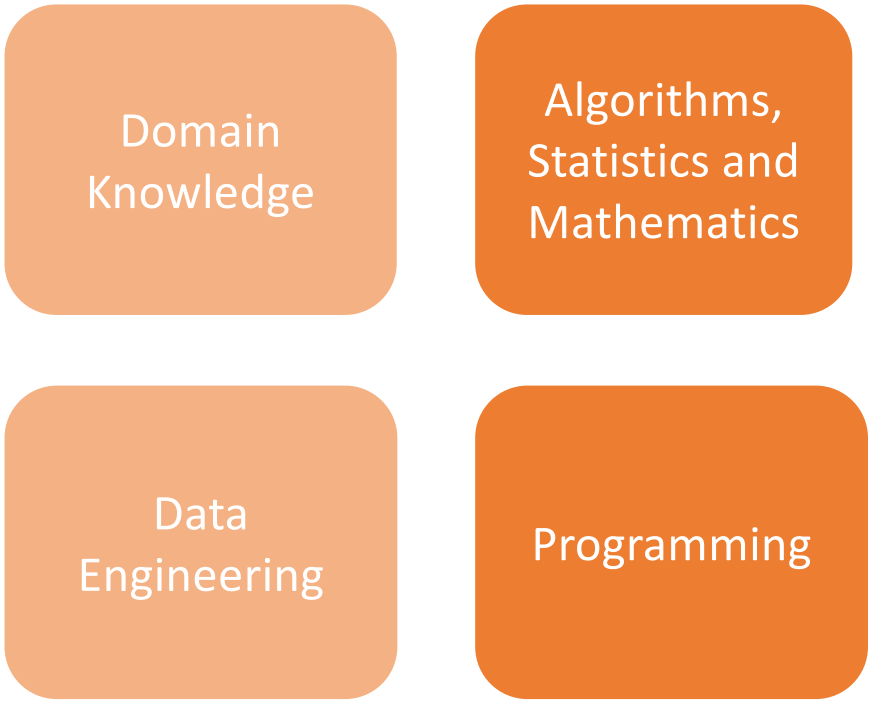
https://www.anaconda.com/products/distribution

colab

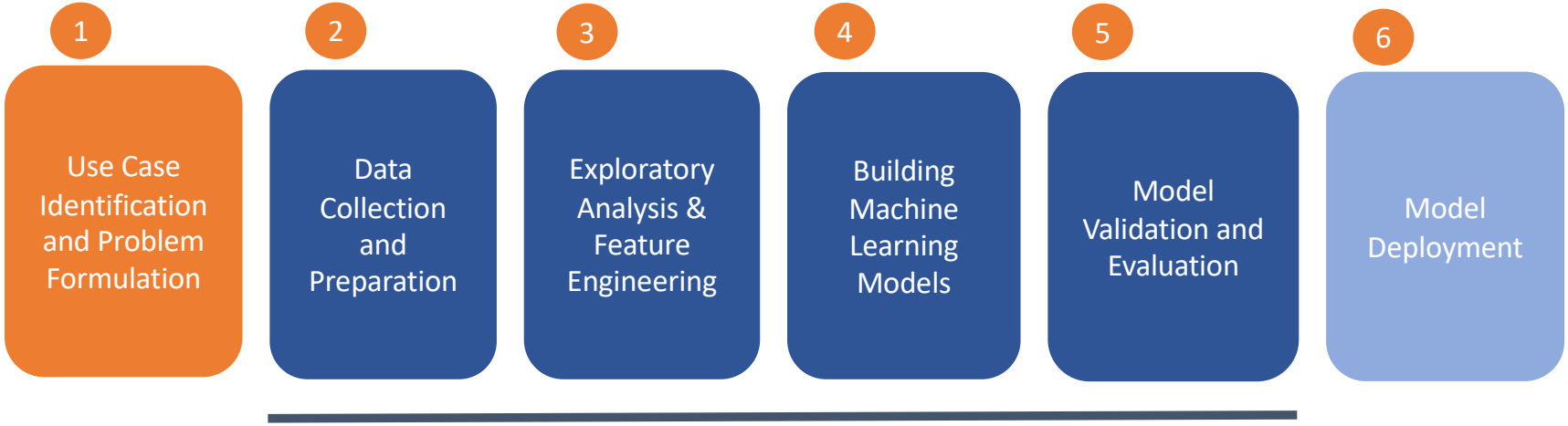
Goole Colaboratory is a hosted Jupyter notebook environment that is free to use and requires no setup.

https://colab.research.google.com/

Key Skills required for Machine Learning



ML Lifecycle



Iterative Steps in Model Development

Exploratory Data Analysis

Building a portal for sale of used cars

- Listing of cars
- Search cars by attributes
- Quotation by sellers
- Negotiation by buyer and seller
- Final transaction

Sellers

Buyers

Factors impacting the value of an used car

Sellers

Problem of underquoting, and overquoting the sale price. This may lead to loss or customer dissatisfaction.

Buyers

Problem: How to estimate the sale price of an used car based on the information of the car?

Can be an value added feature provided by the site to its customers.

Question: List a possible set of factors that may influence the resale value of an used car?

Factors impacting the value of an used car

Possible factors

- Make and model
- Age of the car
- Service Records
- Fuel Type
- Kilometer Driven
- Condition of the car
- Color of the Car
- Mileage

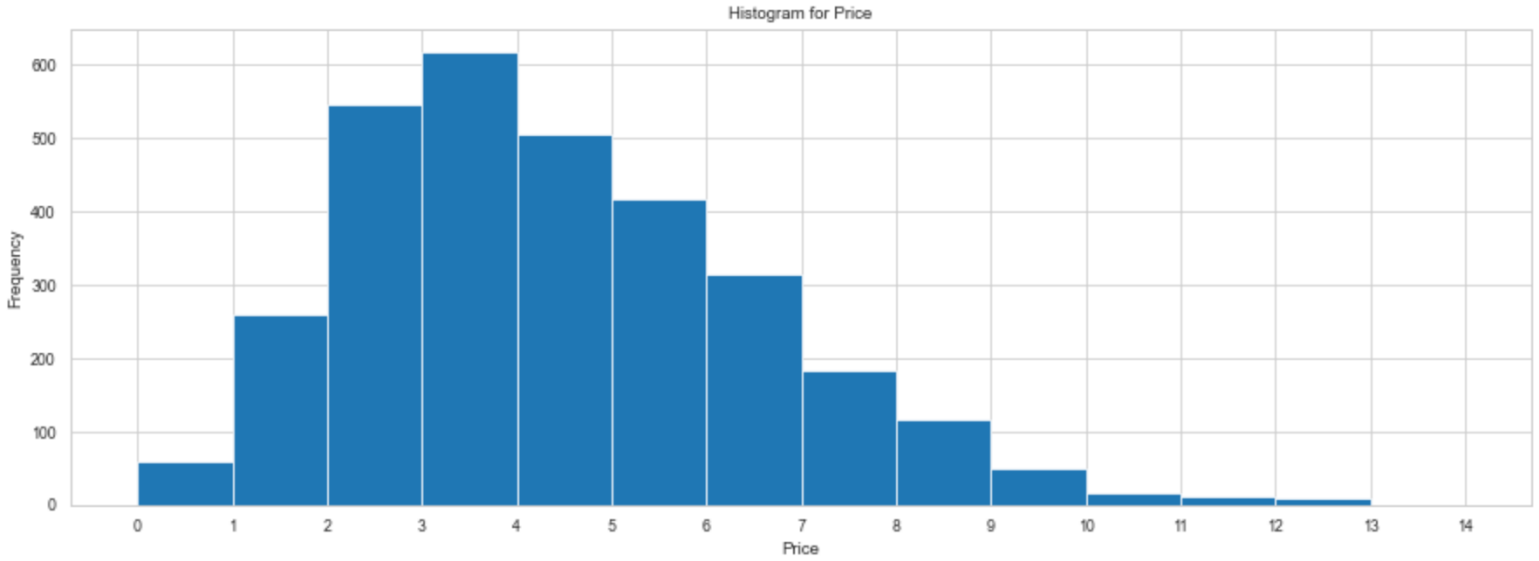
In used car showrooms, an indicative selling price is typically generated by an sales agent who has on-the-ground experience of the used car market as he/she interacts with buyers and understands the requirements and demand.

- Domain Experts
- Interaction with business is very important at the stage of problem definition and identification of factors

Dataset

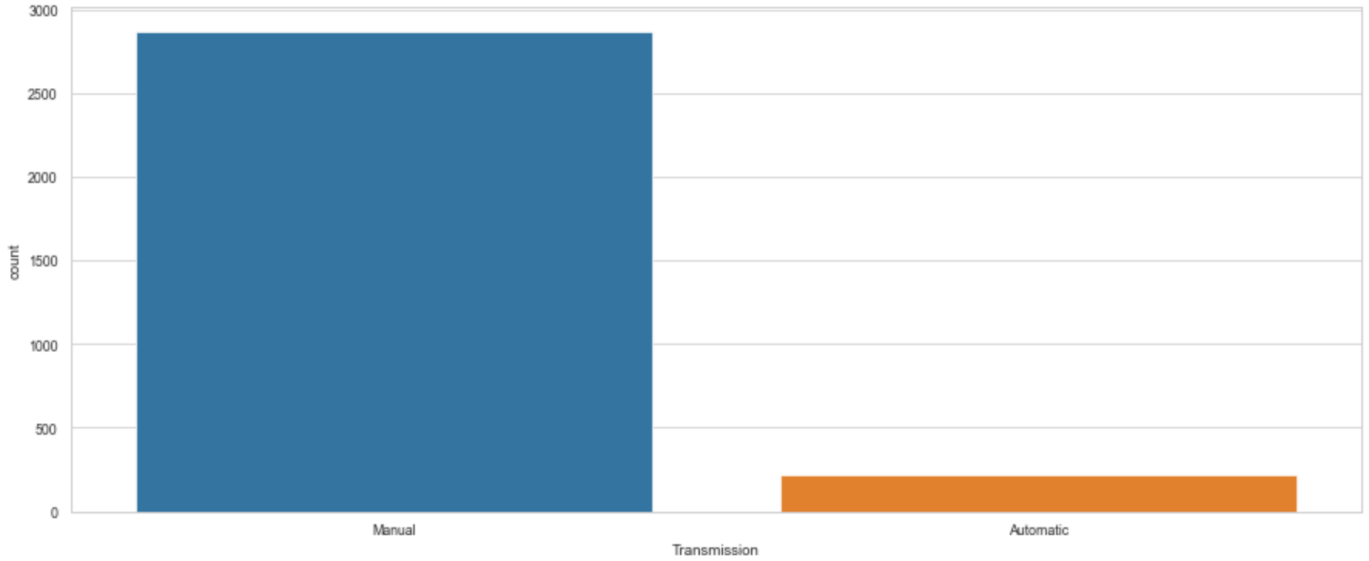
Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
Maruti Swift VXI	Mumbai	2016	23555	Petrol	Manual	First	22.0 kmpl	1197 CC	81.80 bhp	5.0	7 Lakh	5.10
Maruti Alto 800 LXI	Jaipur	2015	25000	Petrol	Manual	First	22.74 kmpl	796 CC	47.3 bhp	5.0	NaN	2.75
Hyundai Santro Xing GLS	Bangalore	2008	46000	Petrol	Manual	Second	17.92 kmpl	1086 CC	62.1 bhp	5.0	NaN	2.22
Maruti Wagon R LXI BS IV	Delhi	2015	13008	Petrol	Manual	First	20.51 kmpl	998 CC	67.04 bhp	5.0	NaN	3.50
Hyundai Elite i20 Petrol Sportz	Kochi	2018	14223	Petrol	Manual	First	18.6 kmpl	1197 CC	81.86 bhp	5.0	NaN	7.32
Honda Jazz 1.2 V AT i VTEC Privilege	Pune	2016	21000	Petrol	Automatic	First	19.0 kmpl	1199 CC	88.7 bhp	5.0	NaN	6.35
Hyundai i20 1.4 Asta (AT)	Chennai	2009	72000	Petrol	Automatic	Third	15.0 kmpl	1396 CC	100 bhp	5.0	NaN	3.25
Honda Amaze S Petrol	Kolkata	2013	32576	Petrol	Manual	First	19.5 kmpl	1199 CC	88.76 bhp	5.0	7.36 Lakh	3.15
Honda Amaze SX i-VTEC	Coimbatore	2014	28246	Petrol	Manual	Second	17.8 kmpl	1198 CC	86.7 bhp	5.0	NaN	4.94
Maruti Swift Dzire VDi	Hyderabad	2014	123900	Diesel	Manual	First	19.3 kmpl	1248 CC	73.9 bhp	5.0	NaN	5.40

Histogram



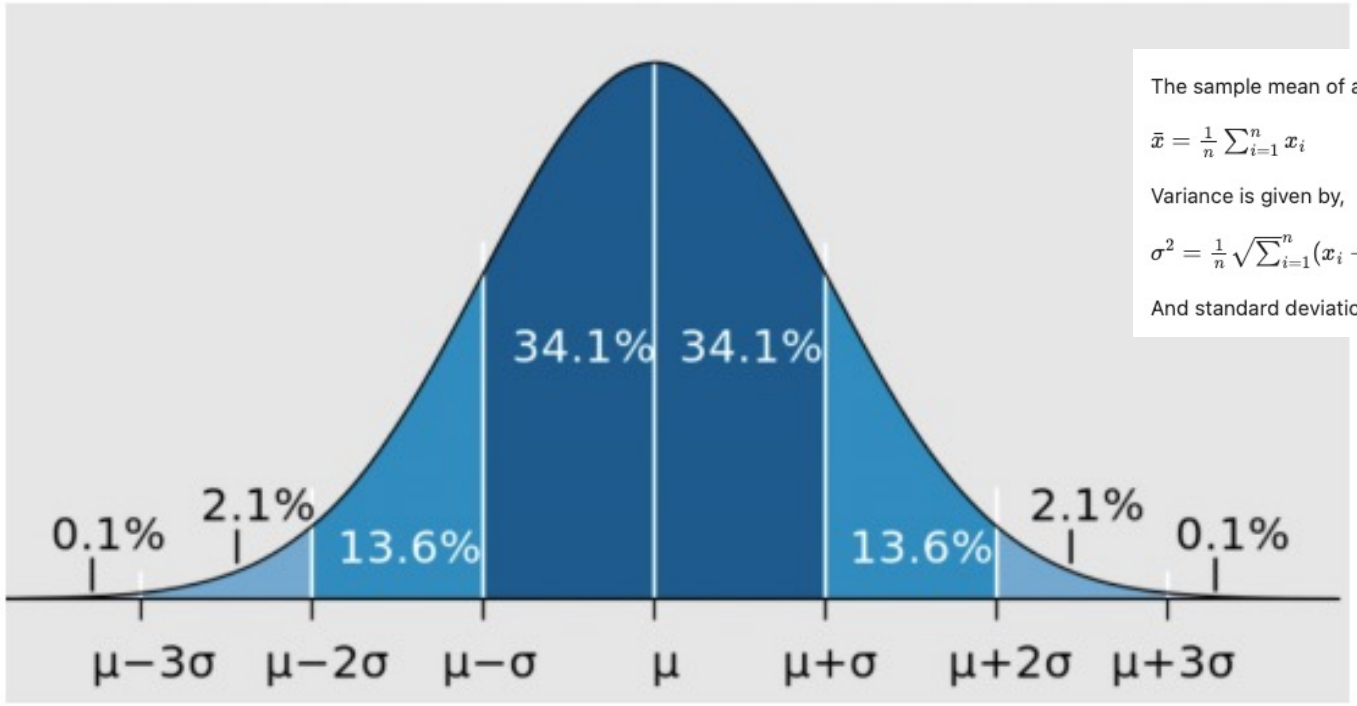

```
plt.figure(figsize=(15,5))
hist_data = plt.hist(cars_df['Price'], bins=list(range(0, 15, 1)));
```

Bar Plot




```
plt.figure(figsize=(15, 6))
sn.countplot(data = cars_df,
             x = 'Transmission');
```

Normal Distribution



The sample mean of a normal distribution is given by,

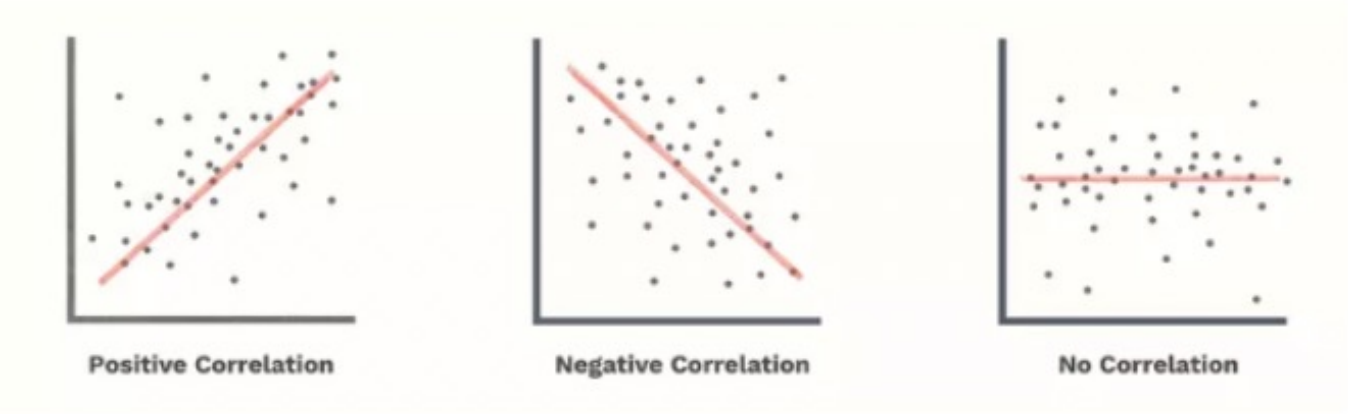
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variance is given by,

$\sigma^2 = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$

And standard deviation is square root of variance and is denoted by σ.

Correlation



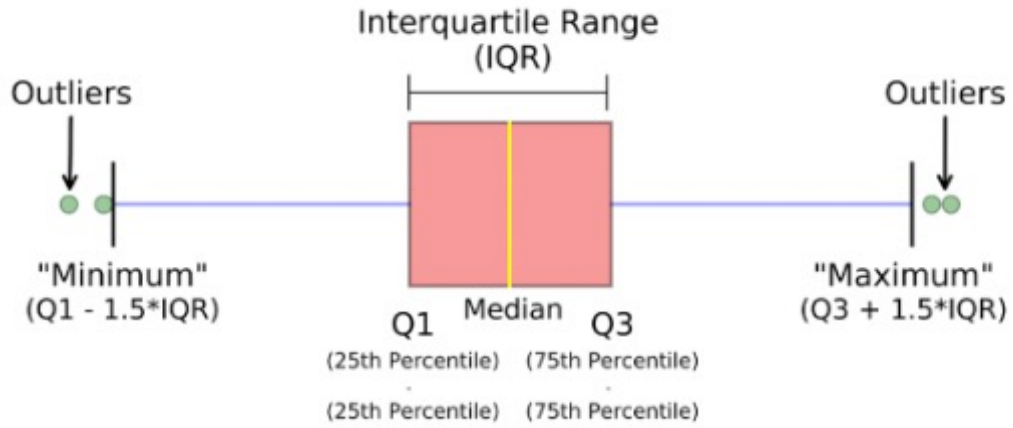
Correlation is given by:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y}$$

This is also known as **Pearson Correlation**.

- | r | < 0.25 - No relationship
- 0.25 < | r | < 0.5 - Weak relationship
- 0.5 < | r | < 0.75 - Moderate relationship
- | r | > 0.75 - Strong relationship

Finding outliers using Box Plot

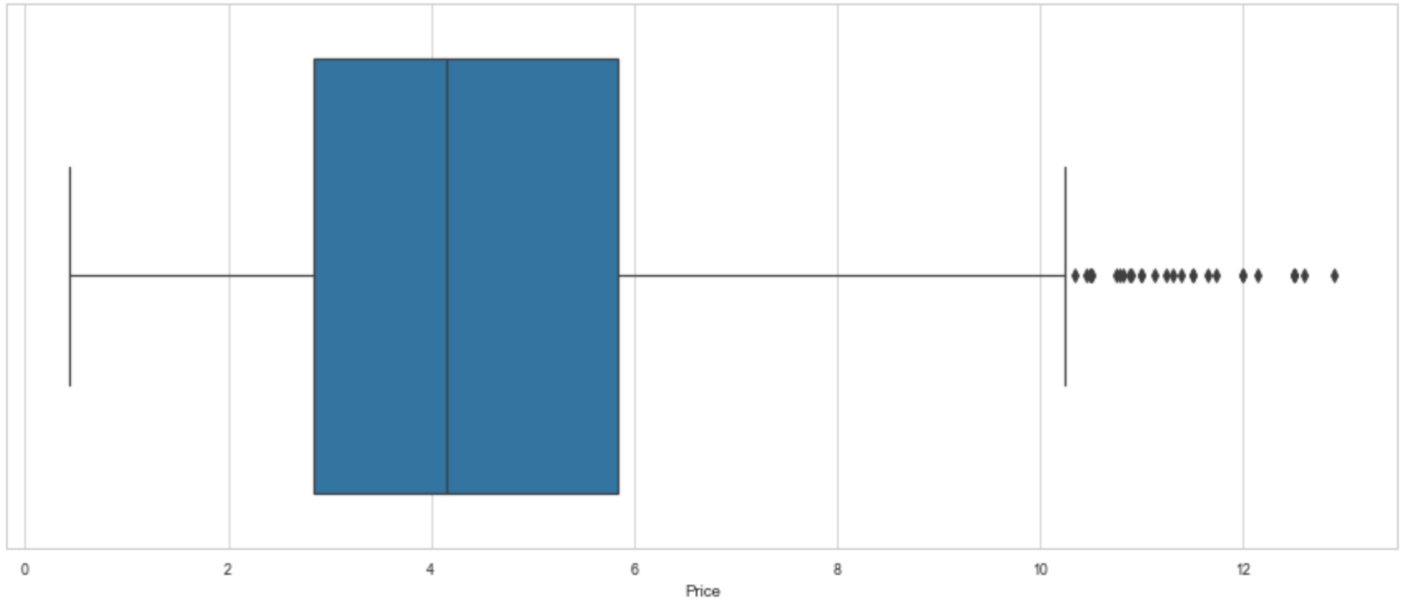


IQR = Q3 – Q1

Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

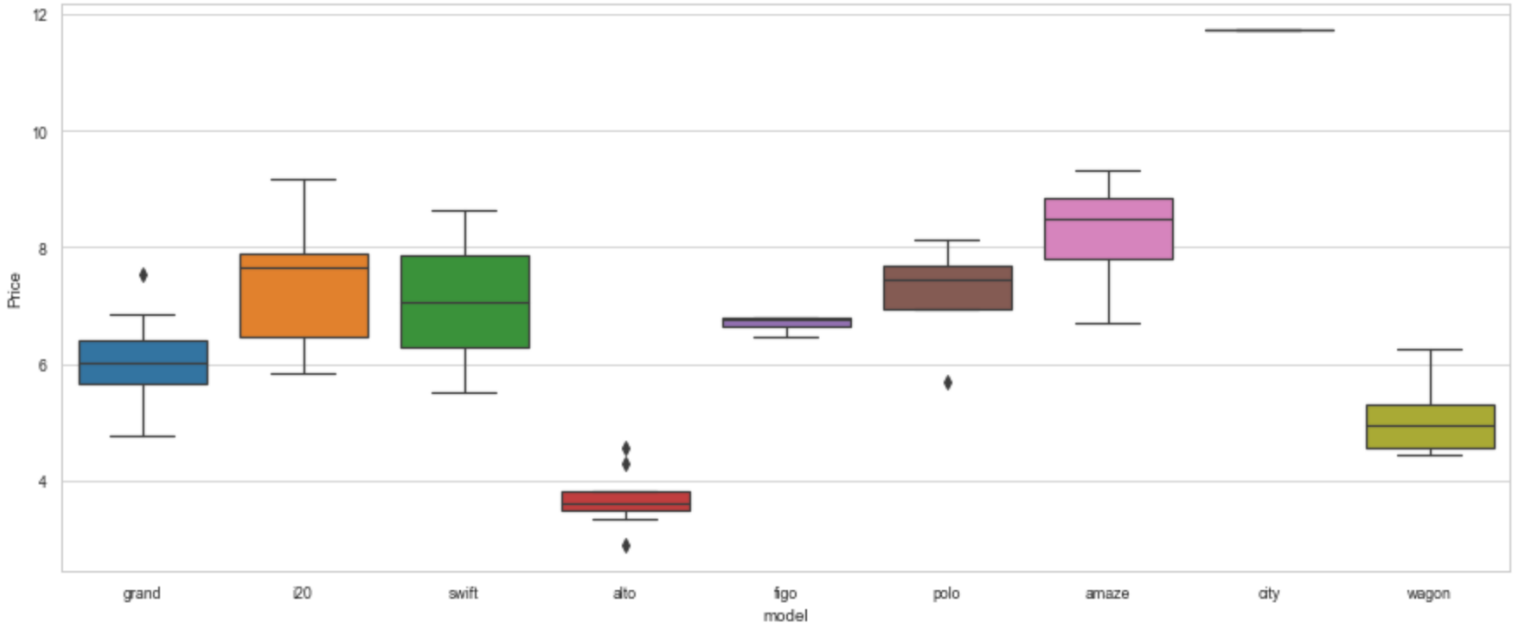
- Calculate the interquartile range for the data.
- Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
- Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
- Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Outlier Analysis



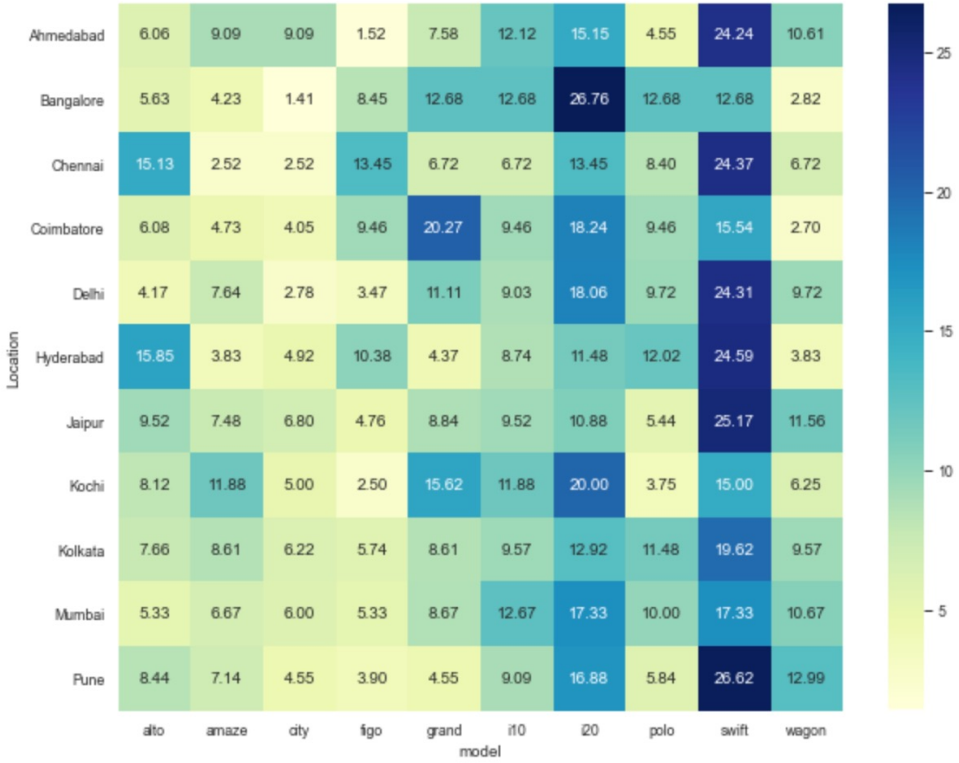

```
plt.figure(figsize=(15,6))
boxp = sn.boxplot(cars_df['Price']);
```

Bivariate: Numerical vs Categorical




```
plt.figure(figsize=(15, 6))
sn.boxplot(data = top_10_models_df,
           x = 'model',
           y = 'Price');
```

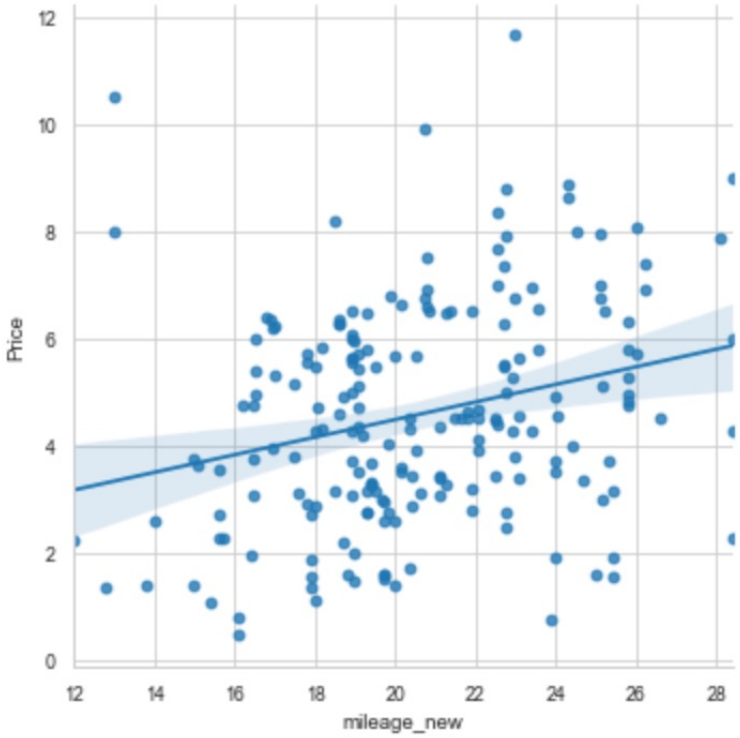
Bivariate: Categorical Variables




```
models_ct = pd.crosstab(top_10_models_df.Location,
                        top_10_models_df.model,
                        normalize = 'index') * 100
```

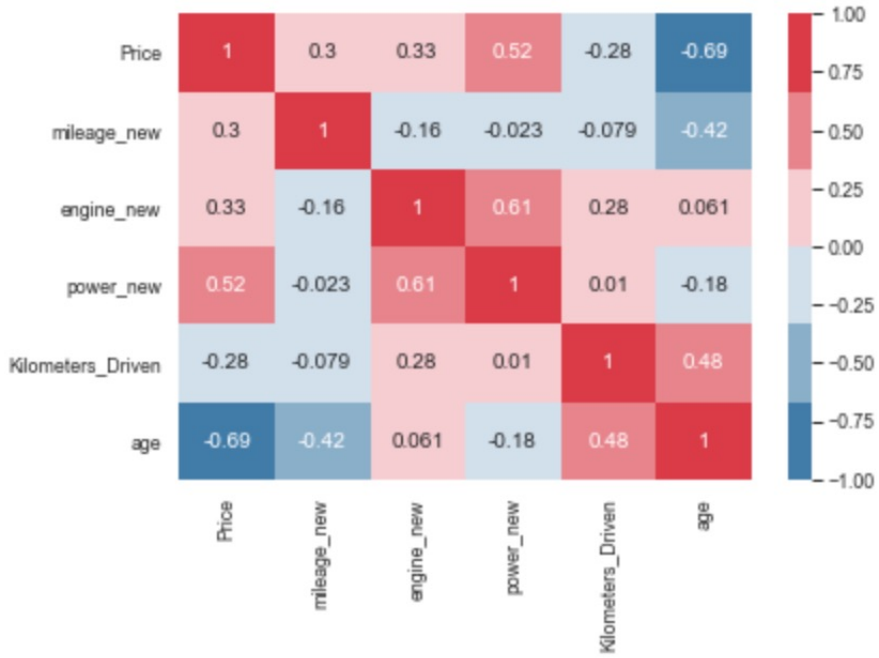
```
plt.figure(figsize=(10, 8))
sn.heatmap(models_ct, annot=True, fmt = "0.2f", cmap="YlGnBu");
```

Bivariate: Two Numerical Variables




```
sn.lmplot(data = cars_df.sample(200),
          x = 'mileage_new',
          y = 'Price');
```

Heatmap




```
sn.heatmap(corr_mat,
           annot=True,
           vmin = -1.0,
           vmax = 1.0,
           cmap = sn.diverging_palette(240, 10));
```

Exploratory Data Analysis

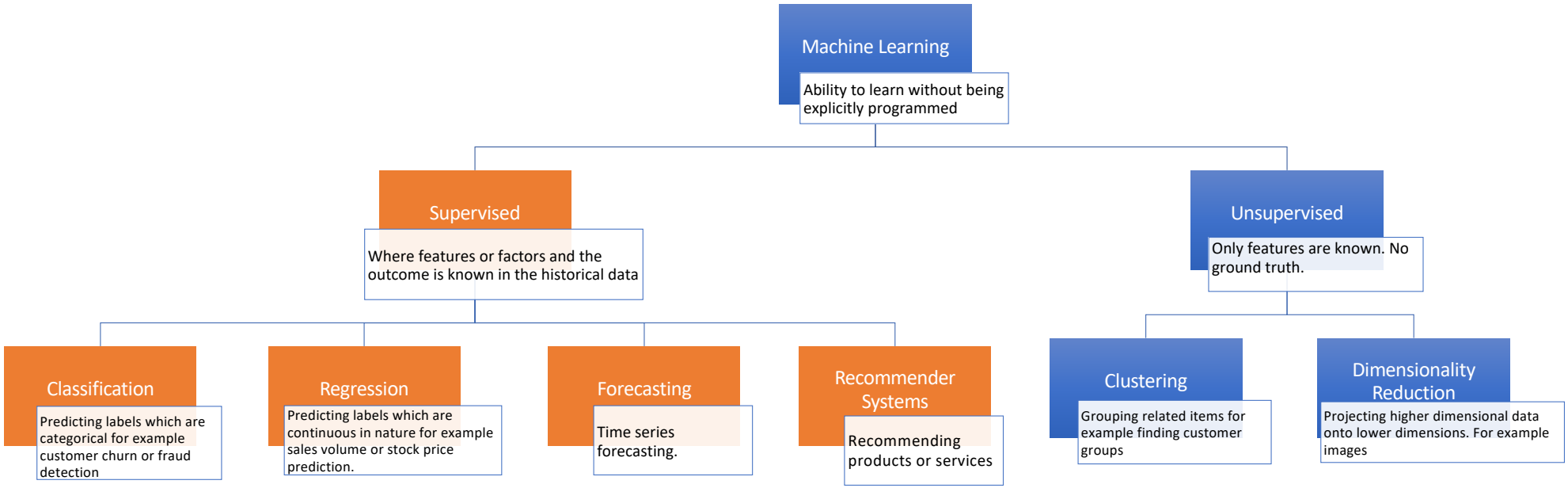
- Understand Distribution of the variables
  - Make use of charts
- Is the data representative?
- Are there any outliers?
- Are there any missing values
  - What is the volume of missing values?
  - Can it be imputed?
- Do we need to transform the variables?
  - Unit transformation?
  - Scaling?

Rules for Plotting

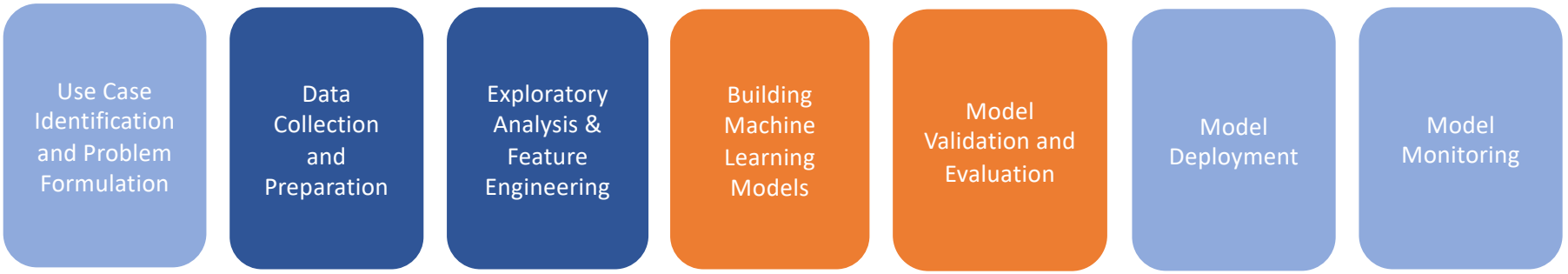
- Single Variable (Univariate Analysis)
  - Continuous -> Histogram, boxplot, distribution plot
  - Categorical -> Count Plot/Bar Plot
- Two Variables (Bivariate Analysis)
  - Continuous + Categorical -> Box plot, Overlapped Distribution Plot
  - Continuous + Continuous -> Scatter Plot, heatmap
  - Categorical + Categorical -> Bar Plot / Count Plot, heatmap

What are different problems that you can solve with ML?

Machine Learning Algorithms



ML Lifecycle



First Model Development Iterative Steps

Continuous Model Update Iterative Steps

Examples of Machine Learning Problems

- What will be the price of a stock in next 3 months given its past performances and the current market outlook?
- What is the estimated demand for a specific product (volume of sale in terms of number of units) in the next quarter given the market conditions and competition?
- What is the likelihood of a customer churning in next 3 months given his/her purchase patterns in the last few months?

Examples of Machine Learning Problems

- What is the likelihood of an employee leaving an organization in next 6 months given his/her performance, behavior and skill demand in the market?
- How to cluster customers together based on their demographics and behavior so that appropriate products or promotions can be targeted to them?
- Which products are customers buying together, which can become candidate for cross selling and up selling?

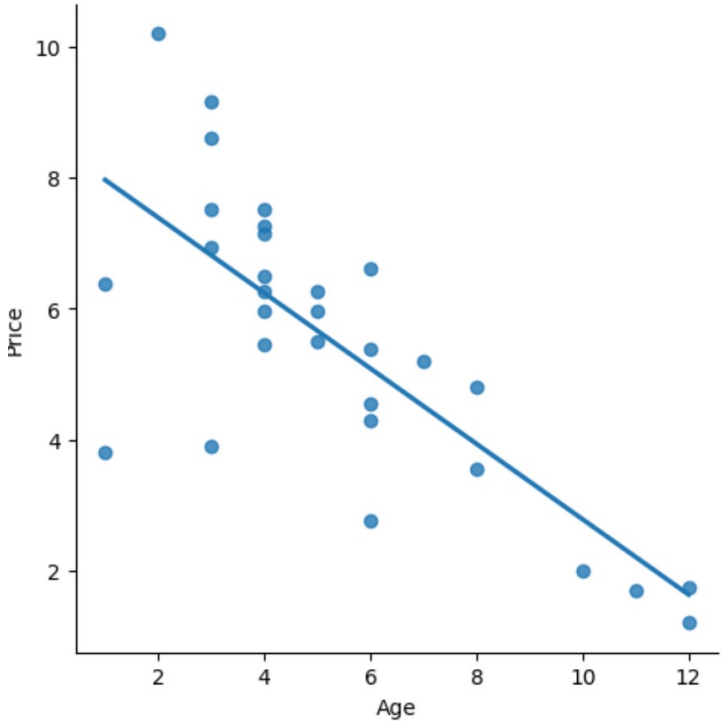
Examples of Machine Learning Problems

- Identify similar products (movies, books etc.) to recommend to customers based on their preferences shown in the past.
- How to identify anomalies in the systems (machines or hardware) so that preventive maintenance can be scheduled rather than periodic maintenance to avoid catastrophic failures?

How to estimate the sale price of an used car?

Regression

Linear Regression



Simple linear regression is given by,

$$\hat{Y} = \beta_0 + \beta_1 X$$

- β_0 and β_1 are the regression coefficients
- $\hat{Y}$ is the predicted value of Y.

So, the error (Mean Squared Error) is:

$$mse = \frac{1}{N} \sum_{i=1}^{n} (Y_i - (\hat{Y}))^2$$

or

$$mse = \frac{1}{N} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Regression: Loss Function

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$\mathbf{MSE}$ = mean squared error

n = number of data points

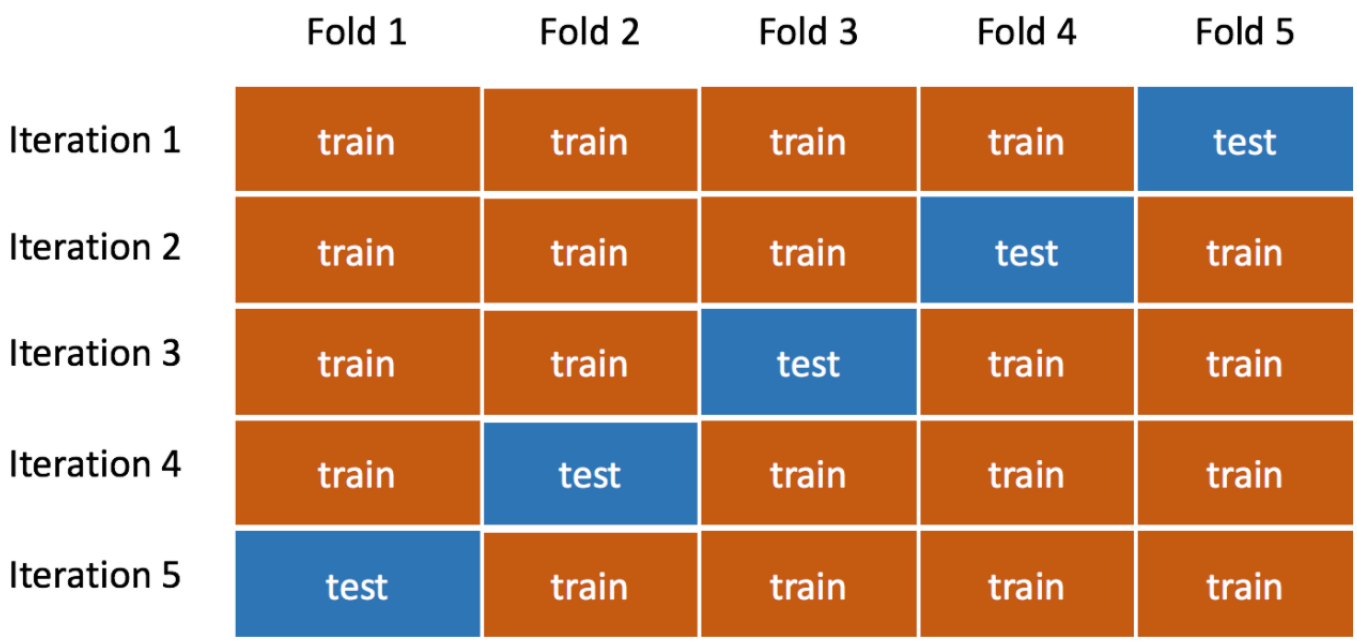
Y_i = observed values

$\hat{Y}_i$ = predicted values

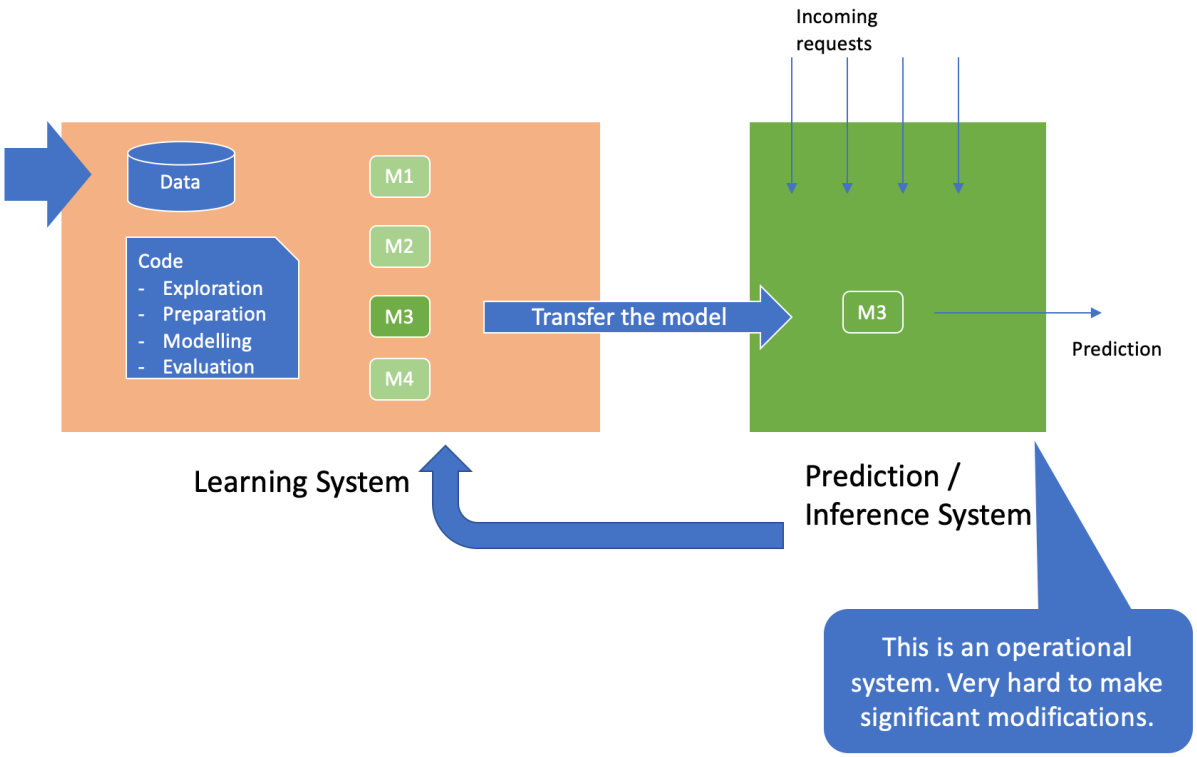
R Squared

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total Variance}}$$

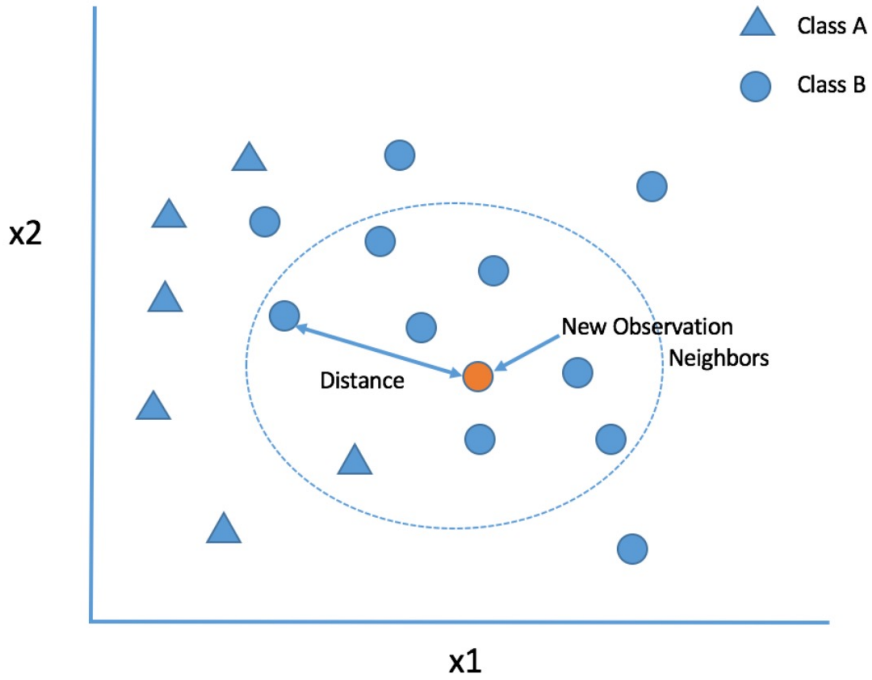
K-Fold Cross Validation



ML Systems


Incoming
requests
Data
Code
- Exploration
- Preparation
- Modelling
- Evaluation
M1
M2
M3
M4
Transfer the model
M3
Prediction
Learning System
Prediction /
Inference System
This is an operational system. Very hard to make significant modifications.

KNN

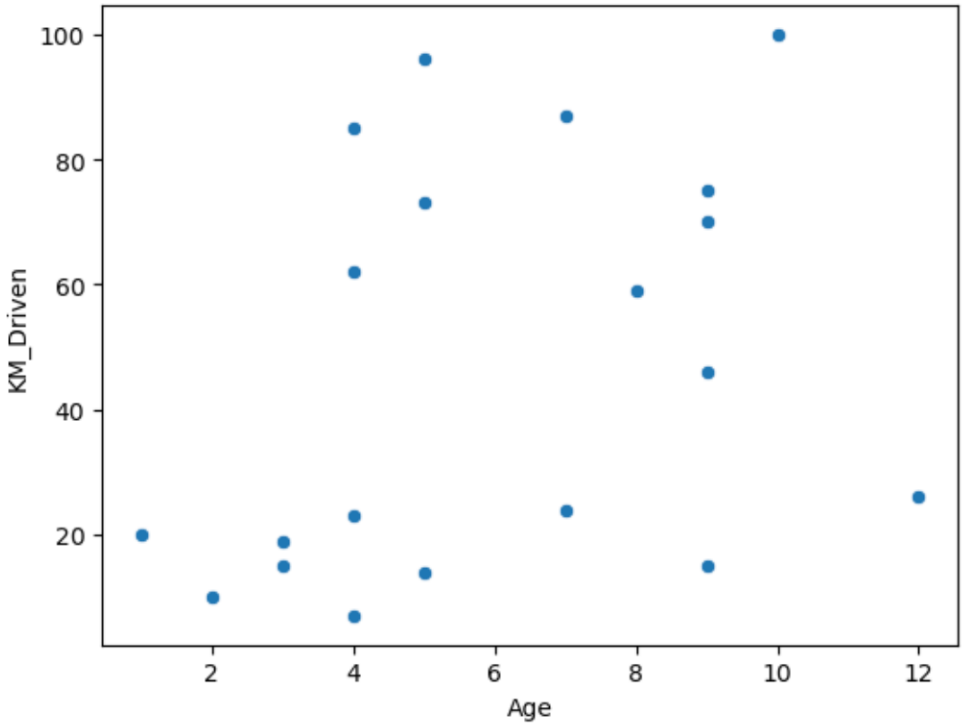


Euclidean Distance

$$D(X1, X2) = \sqrt{\sum_{i=1}^{n} (x_{i1} - x_{i2})^2}$$

Where, X_1 and X_2 are two data points, there are n attributes and x_i is i^{th} attribute of each data points.

Euclidean Distance



$$dist_{xy} = \sqrt{(age_x - age_y)^2 + (km_x - km_y)^2}$$

Scaling

$$X_{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Min Max Scaler

In this technique, the minimum value of the feature is scaled to 0 and the maximum value is scaled to 1. All other values are scaled to a value between 0 and 1 based on their relative position to the minimum and maximum values.

$$X_{norm} = \frac{X_i - \mu}{\sigma}$$

Standard Scaler

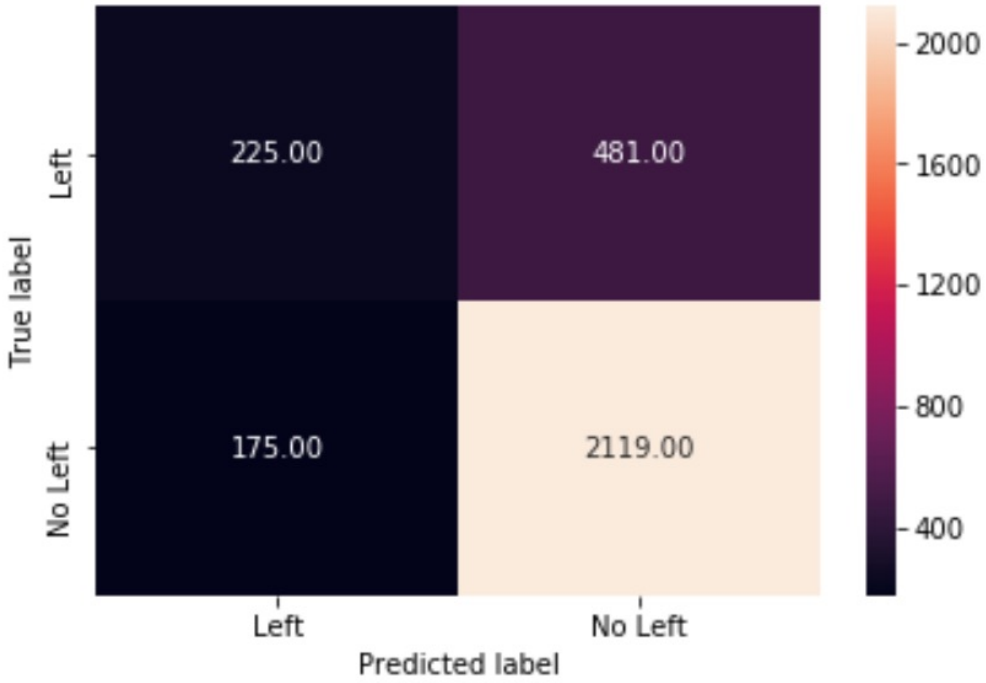
Standard scaling, also known as standardization, is a data preprocessing technique used in machine learning and data science to transform the features of a dataset so that they have a mean of 0 and a standard deviation of 1.

HR Attrition Case Study

We spend huge amount of time and money to hire and nature their employees. It is a huge loss for companies if employees leave, especially the key resources.

Can we identify the attrition risks and and make policy or employee level intervention to retain those employees or do preventive hiring to minimize the impact.

Confusion Matrix



Classification Metrics

Precision is defined as how many are actual positives out of total number of positives identified by the model and is defined as

$$TPR = (\frac{TP}{TP+FP})$$

True Positive Rate (TPR) or Recall or Sensitivity is how many actual positive are properly identified by the model out of total number actual positive in the test set and is defined as

$$TPR = (\frac{TP}{TP+FN})$$

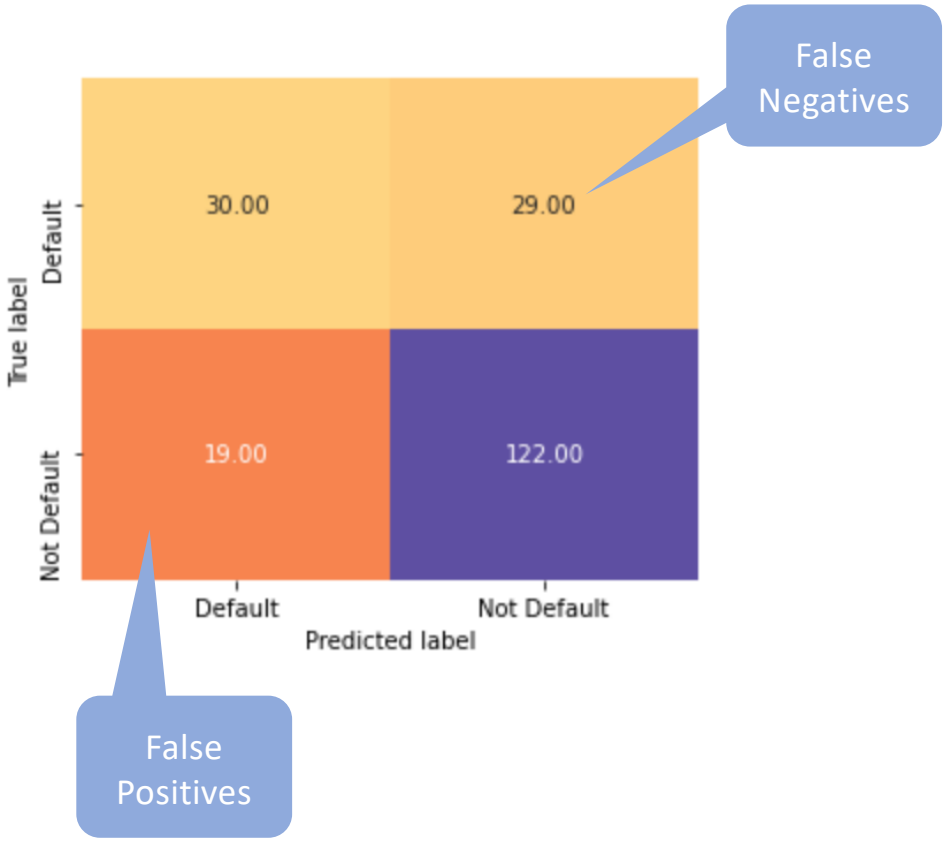
True Negative Rate (TNR) or Specificity is how many are correctly indentified as correct negatives out of all acutal negative present in the test set and is defined as

$$TNR = (\frac{TN}{FP+TN})$$

F-Score (F-Measure) is another measure used in binary logistic regression that combines both precision and recall (harmonic mean of precision and recall) and is given by

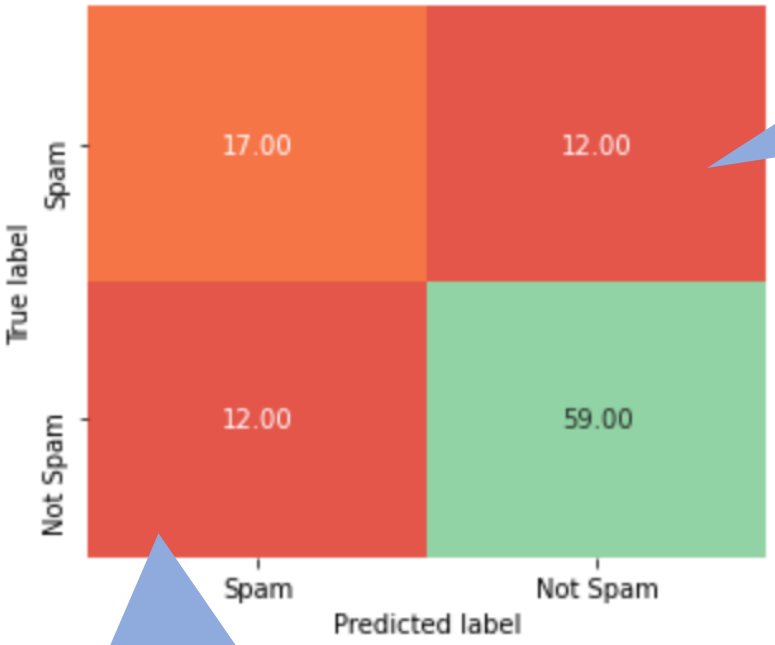
$$F1 - score = (\frac{2xPrecisionxRecall}{Precision+Recall})$$

Cost of Failure: Loan Default



$$\text{Recall} = \frac{TP}{(TP+FN)}$$

Cost of Failure: Spam Detection Model

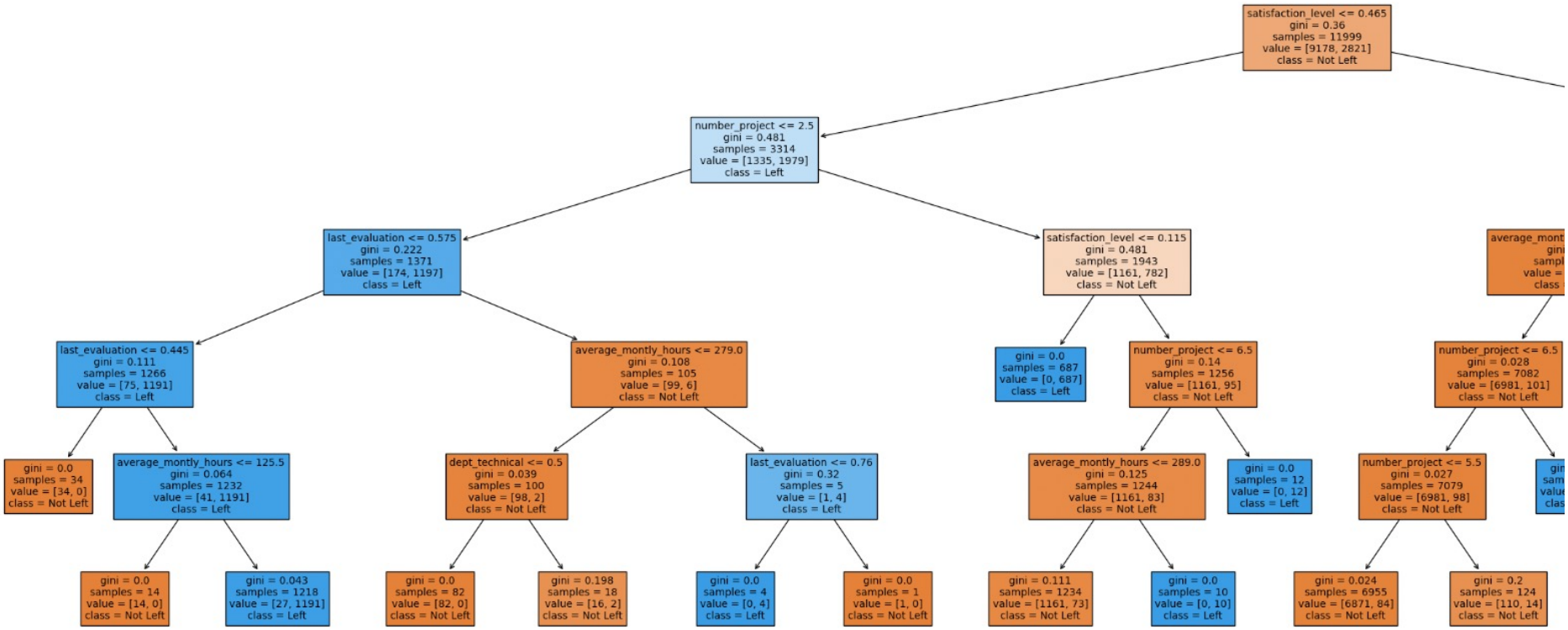


Spam mails that appear in the inbox

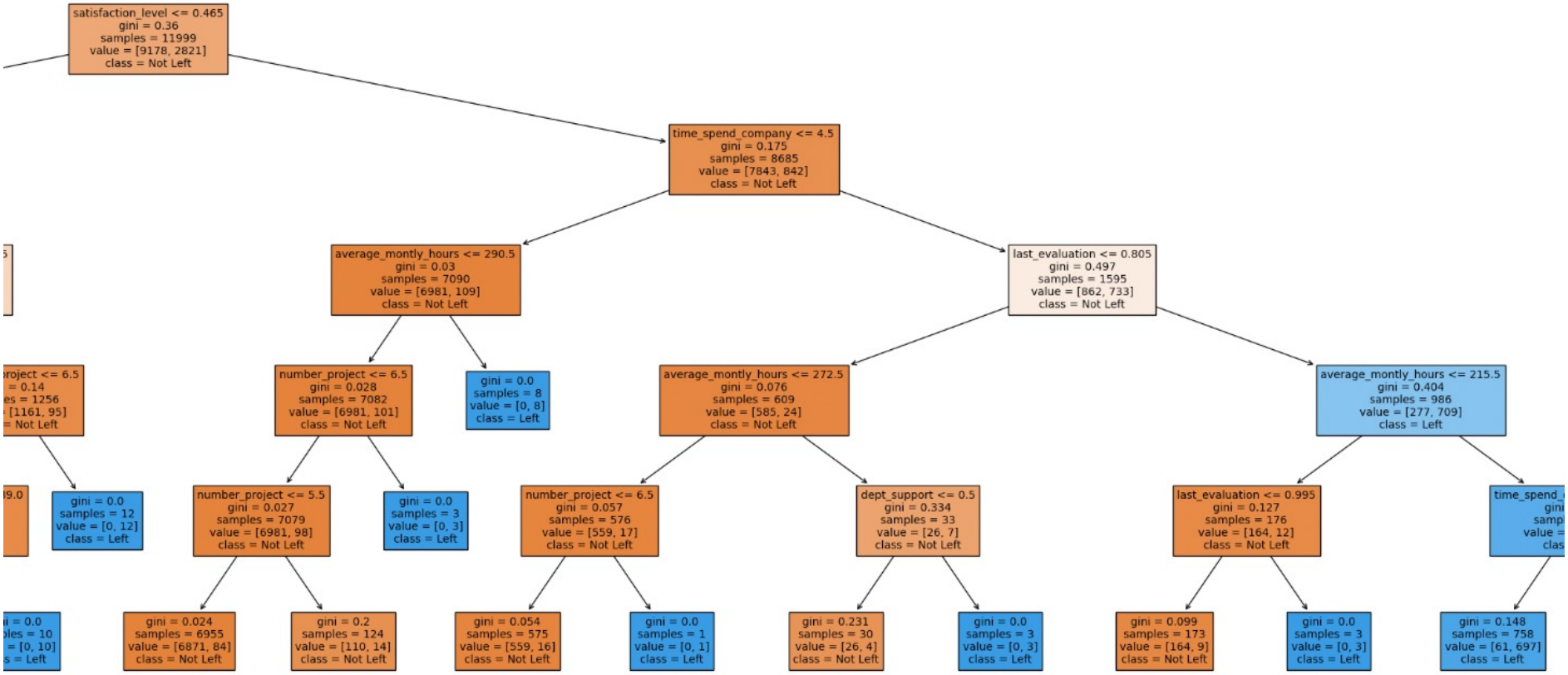
Good mails that appear in the spam box

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Decision Tree Model



Decision Tree Model



Gini & Entropy

$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

$$Entropy = \sum_{i=1}^{n} -p(c_i)log_2(p(c_i))$$

where $p(c_i)$ is the probability/percentage of class c_i in a node.

Feature Importance

features	importance	cumsum
satisfaction_level	0.522042	0.522042
time_spend_company	0.158409	0.680451
last_evaluation	0.150507	0.830958
number_project	0.102832	0.933790
average_montly_hours	0.066210	1.000000
Work_accident	0.000000	1.000000
promotion_last_5years	0.000000	1.000000
salary	0.000000	1.000000

Earnings Manipulation

Earnings manipulation, also known as financial statement manipulation, is the act of altering financial reports to misrepresent a company's performance. It can involve inflating revenue, hiding expenses, or misrepresenting assets.

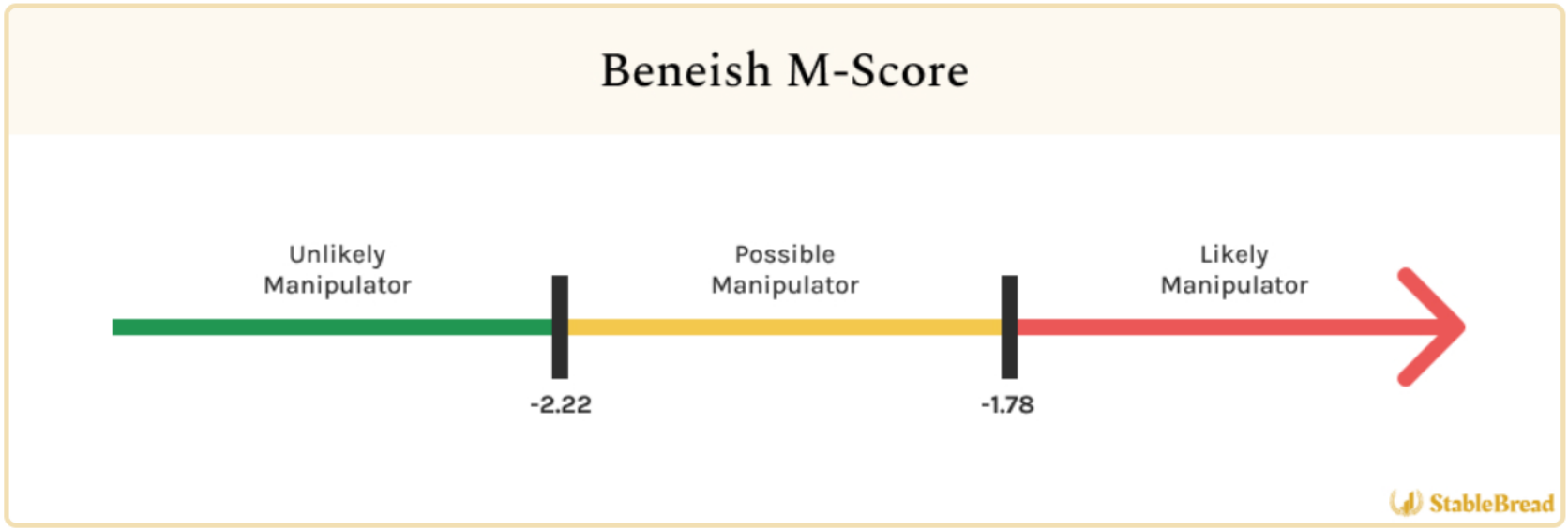
Company	Year	Amount Manipulated	Method Used
Enron	2001	~$63.4 billion	Off-balance-sheet entities, mark-to-market accounting
WorldCom	2002	~$11 billion	Capitalizing expenses instead of recording them as costs
HealthSouth	2003	~$2.7 billion	Inflated revenues through fictitious transactions
Lehman Brothers	2008	~$50 billion	Repo 105 transactions to hide debt

An ML Approach

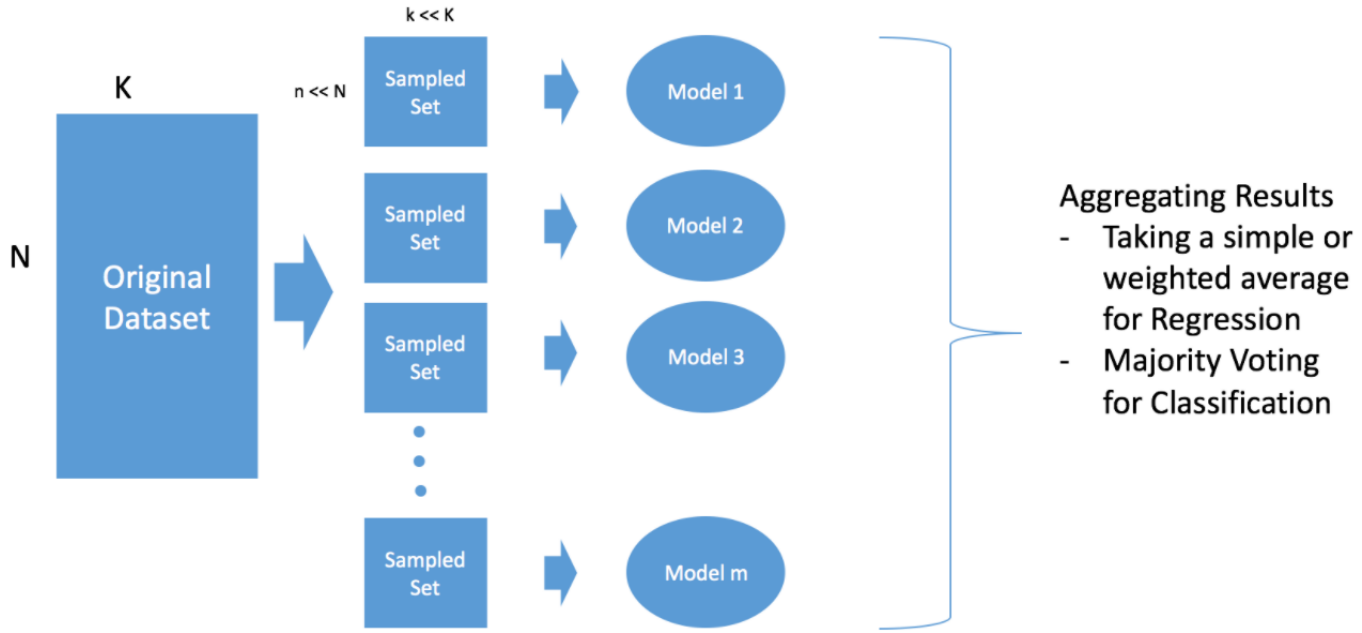
$$P(\text{a balance sheet is manipulated}) = f\left(\right.$$

$$\left.\right)$$

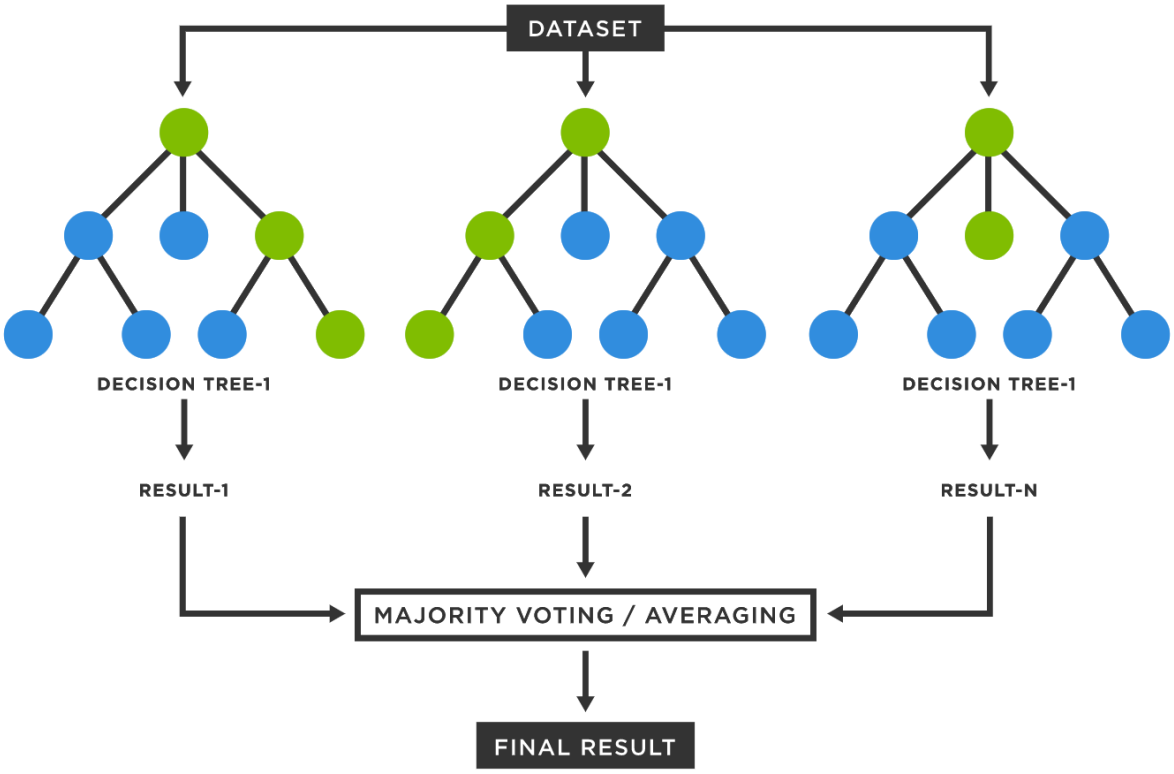
Earnings Manipulation



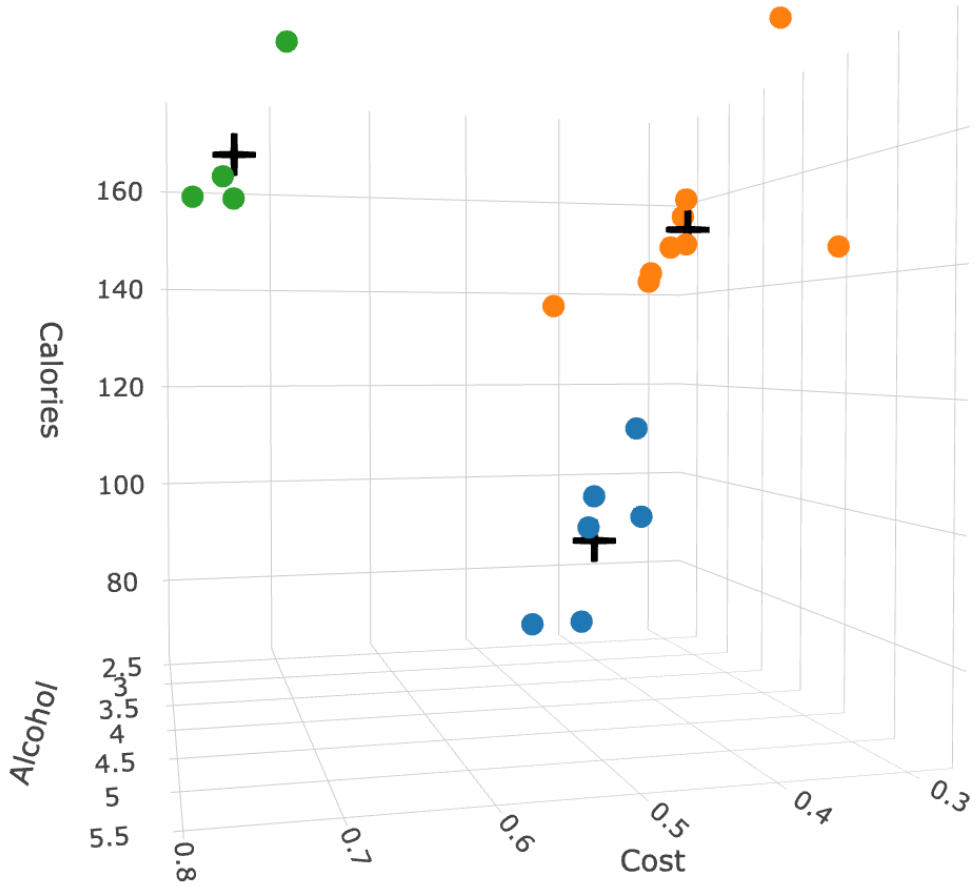
Ensemble


k << K
K
n << N
Sampled Set
Model 1
N
Original Dataset
Sampled Set
Model 2
Sampled Set
Model 3
Sampled Set
Model m
Aggregating Results
- Taking a simple or weighted average for Regression
- Majority Voting for Classification

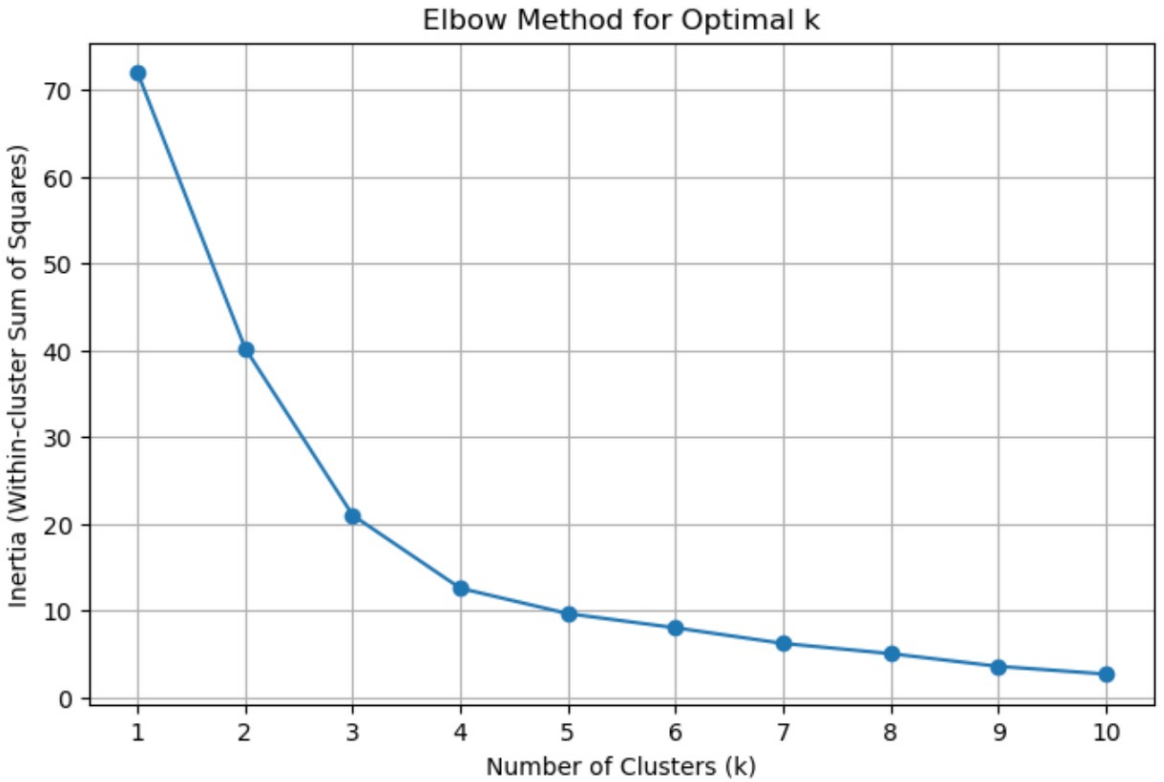
Random Forest



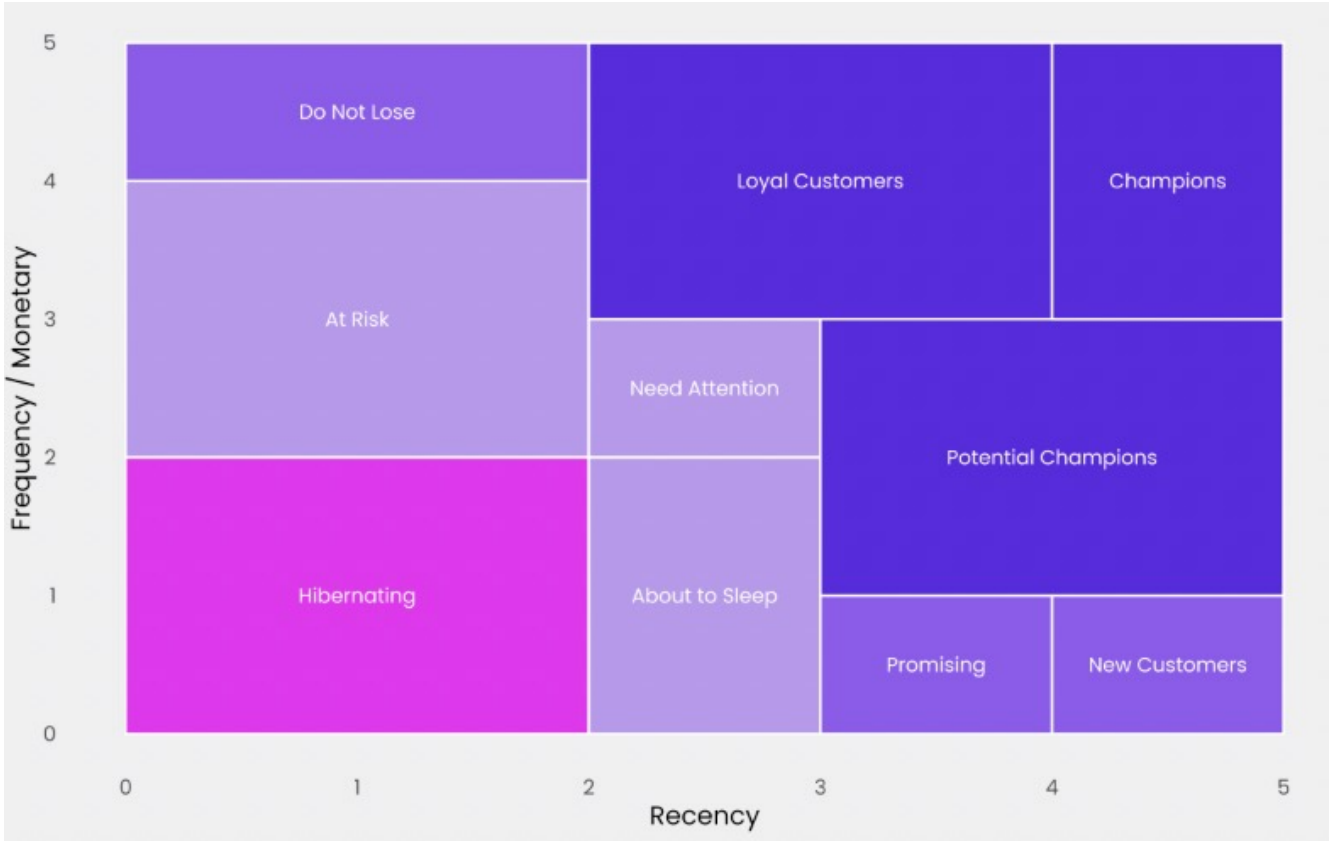
Clustering


160
140
120
100
80
Calories
2.5
3
3.5
4
4.5
5
5.5
Alcohol
0.8
0.7
0.6
0.5
0.4
0.3
Cost

Elbow Method



RFM Segmentation



https://ctrldigital.com/posts/rfm-segmentation-using-bigquery/