Predictive Modelling

Process of Predictive Modelling → future →

1. **Creating the model :** Software solutions allows you to create a model to run one or more algorithms on the data set →

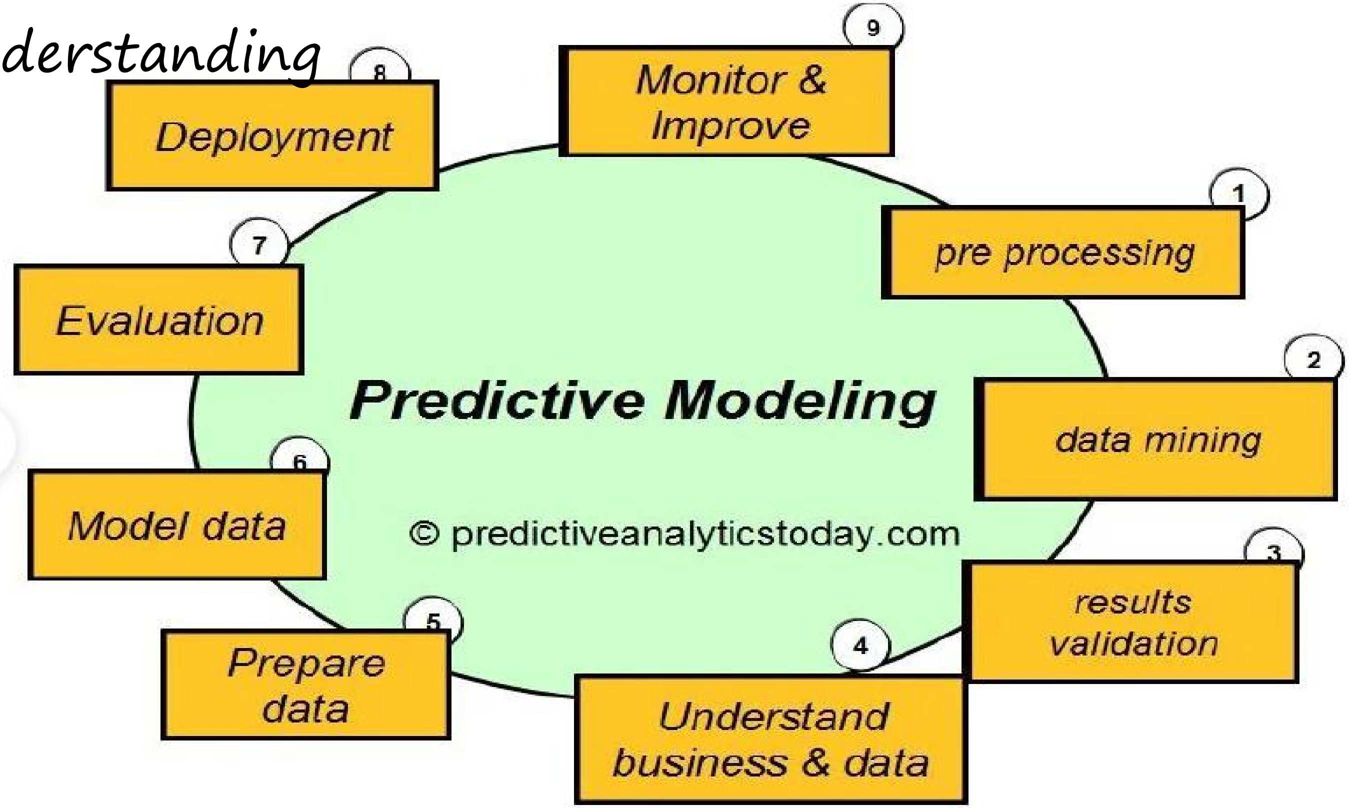
2. **Testing the model:** Test the model on the data set. In some scenarios, the testing is done on past data to see how best the model predicts

3. **Validating the model :** Validate the model run results using visualization tools and business data understanding →

4. **Evaluating the model :** Evaluating the best model from the models used and choosing the model right fitted for the data →

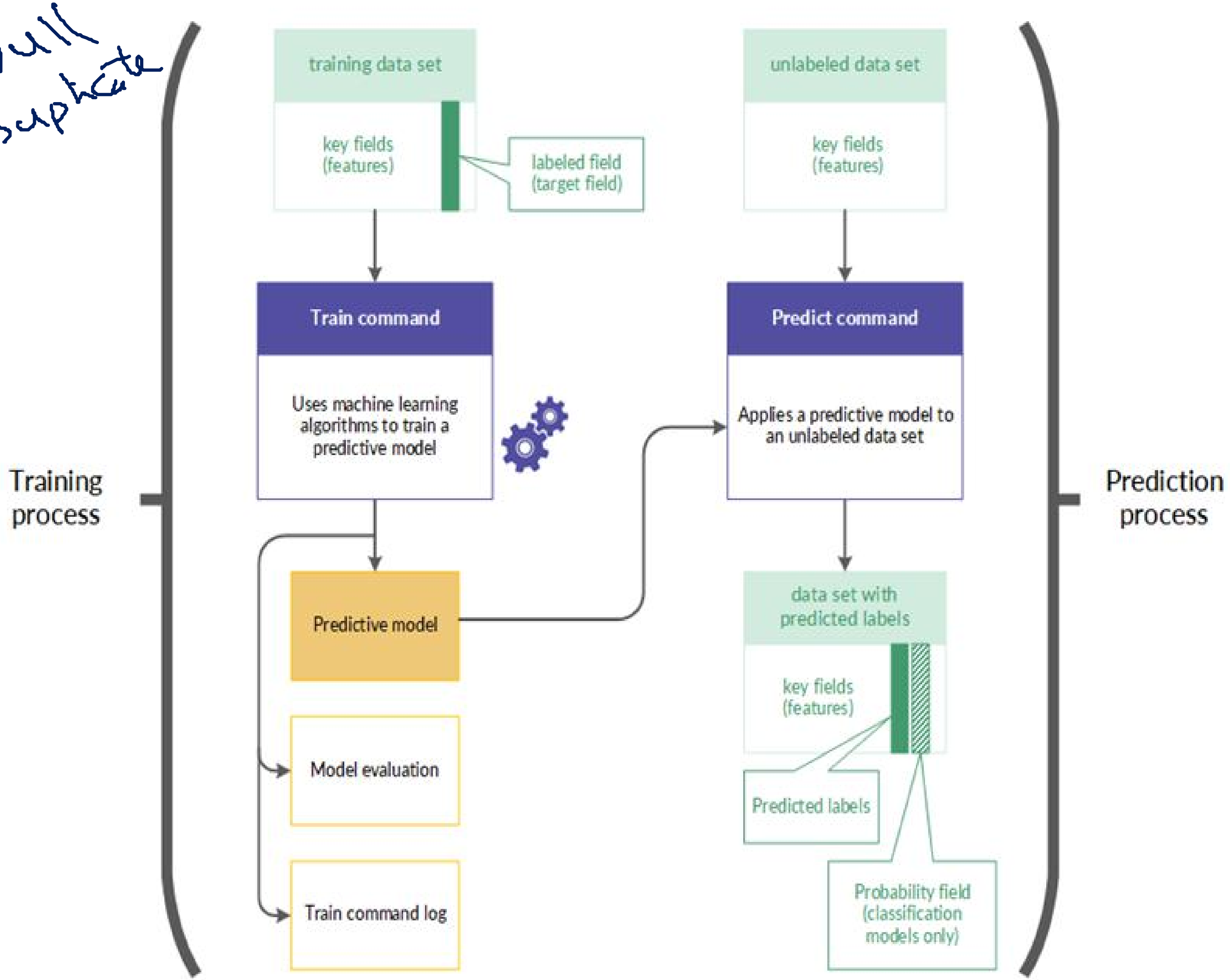
Predictive Modeling Process

- The process involve running one or more algorithms on the data set where prediction is going to be carried out.
- This is an iterative processing and often involves **training the model, using multiple models** on the same data set and finally arriving on the best model based on the business data understanding



Predictive Process

1. Data Gathering and Cleansing — Null Duplicate
2. Data Analysis/Transformation
3. Building a Predictive Model
4. Inferences/Evaluation



Types of predictive models

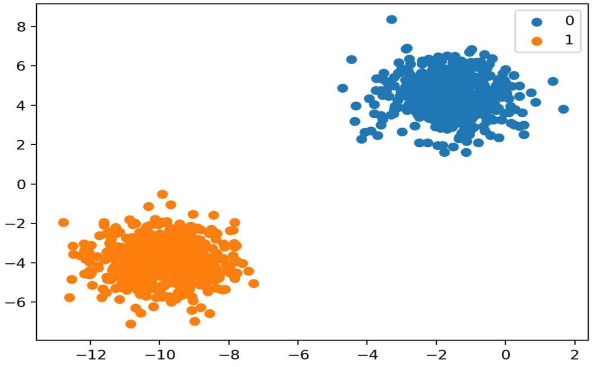
- **1. Classification model:** Considered the simplest model, it categorizes data for simple and direct query response. An example use case would be to answer the question "Is this a fraudulent transaction?"
- Among all the predictive modelling techniques in machine learning, the classification model is one of the widely used techniques. In classification predictive modelling, an input is classified into a specific category where it is treated as a label and its class is predicted.
- In predictive modelling, a general data point is inserted in the software that classifies the input and predicts the class of the output.

Classification model

- Classification refers to a predictive modeling **problem where a class label is predicted for a given example of input data**.
- Ex: Given an example, classify if it is spam or not. Given a handwritten character, classify it as one of the known characters.
- Classification predictive modeling involves assigning a class label to input examples.
- In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

Classification model

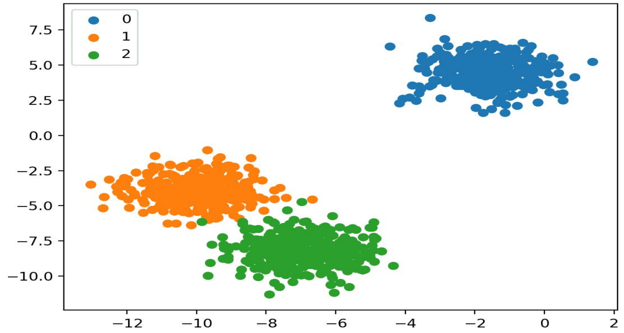
- Binary classification refers to those classification tasks that have two class labels.
- Examples: Email spam detection (spam or not)
- Churn prediction (churn or not). ↔
- Conversion prediction (buy or not).



- binary classification tasks involve one class that is the normal state and another class that is the abnormal state.
- class for the normal state is assigned the class label 0 and the class with the abnormal state is assigned the class label 1.

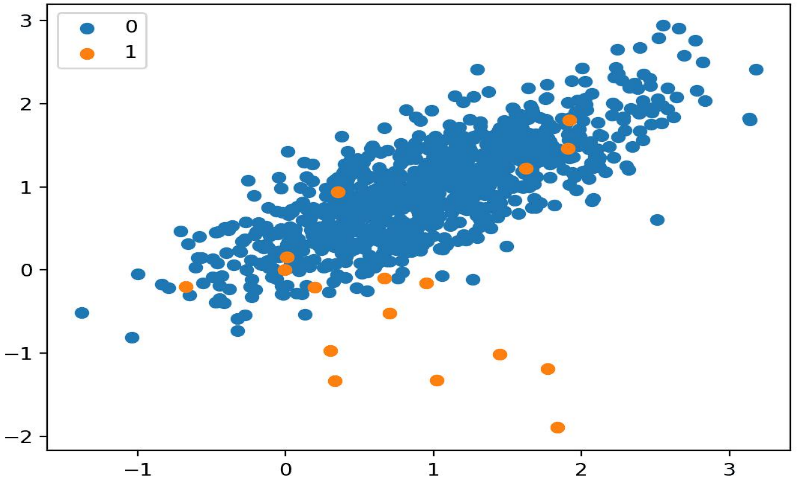
Classification model

- Multi-class classification refers to those classification tasks that have more than two class labels.
- Face classification.
- Plant species classification.
- Optical character recognition.
- number of class labels may be very large on some problems
- Example: a model may predict a photo as belonging to one among thousands or tens of thousands of faces in a face recognition system.
- predicting a sequence of words



Classification model

- Imbalanced Classification: classification tasks where the number of examples in each class is unequally distributed.
- majority of examples in the training dataset belong to the normal class and a minority of examples belong to the abnormal class.
- Example: Fraud detection.
- Outlier detection.
- Medical diagnostic tests.



Types of predictive models

2. Clustering model: Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups

Example: determining credit risk for a loan applicant

Hard Clustering: each customer is put into one group out of the 10 groups.

Soft Clustering: each costumer is assigned a probability to be in either of 10 clusters of the retail store.

Types of predictive models

- 3. **Forecast model:** numerical value based on learning from historical data. metric value prediction. model can be applied wherever historical numerical data is available.
- forecast model also considers multiple input parameters.
- Example: Order requirement of the restaurant for next week.
- prediction of the support calls in a call center per hour.

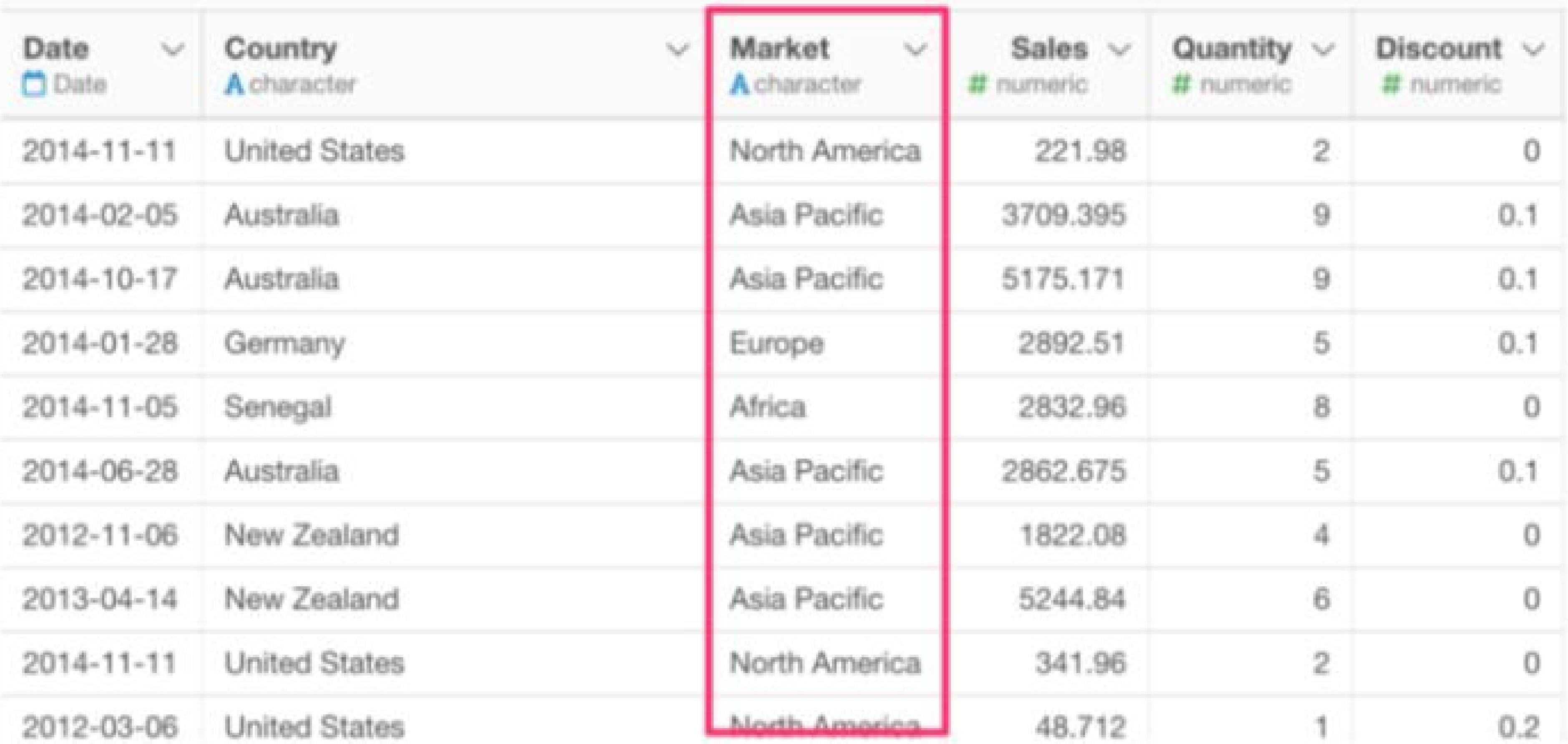
Types of predictive models

- 4. **Outliers model:** outliers model works with anomalous data entries within a dataset. analyzing abnormal or outlying data points.
- Example: Recording a spike in support calls, which could indicate a product failure that might lead to a recall
- Finding anomalous data within transactions, or in insurance claims, to identify fraud
- useful for predictive analytics in retail and finance.
- the model can assess not only amount, but also location, time, purchase history and the nature of a purchase

Types of predictive models

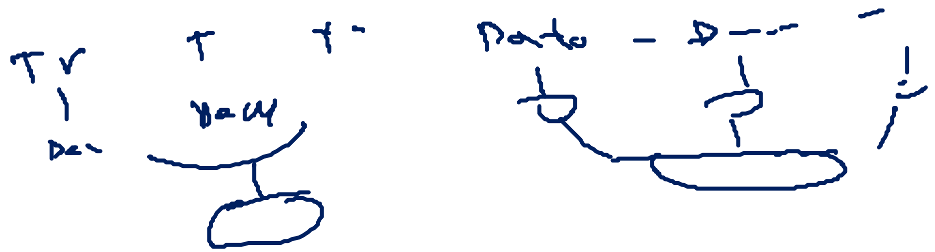
- 5. **Time series model:** evaluates a sequence of data points based on time. The time series model focuses on data where time is the input parameter. different data points (taken from the previous year's data) to develop a numerical metric that will predict trends within a specified period.
- a data set that tracks a sample over time.
- Example: number of stroke patients admitted to the hospital in the last four months
- Weather records, economic indicators

Time series model-Example



Date	Country	Market	Sales	Quantity	Discount
Date	A character	A character	# numeric	# numeric	# numeric
2014-11-11	United States	North America	221.98	2	0
2014-02-05	Australia	Asia Pacific	3709.395	9	0.1
2014-10-17	Australia	Asia Pacific	5175.171	9	0.1
2014-01-28	Germany	Europe	2892.51	5	0.1
2014-11-05	Senegal	Africa	2832.96	8	0
2014-06-28	Australia	Asia Pacific	2862.675	5	0.1
2012-11-06	New Zealand	Asia Pacific	1822.08	4	0
2013-04-14	New Zealand	Asia Pacific	5244.84	6	0
2014-11-11	United States	North America	341.96	2	0
2012-03-06	United States	North America	48.712	1	0.2

Predictive Algorithms



- **1. Random forest** : consisting of many decisions trees. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- **Greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**
- should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

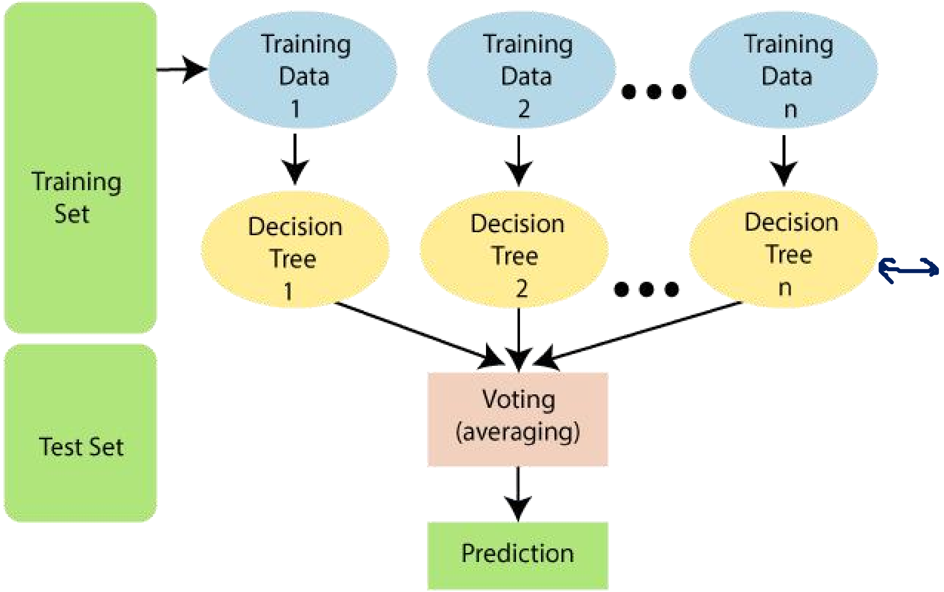
1. Random forest:

It takes less training time as compared to other algorithms.

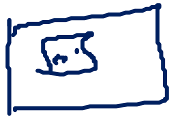
It predicts output with high accuracy, even for the large dataset it runs efficiently.
It is capable of handling large datasets with high dimensionality.

Applications:
Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.

Marketing: Marketing trends can be identified using this algorithm.

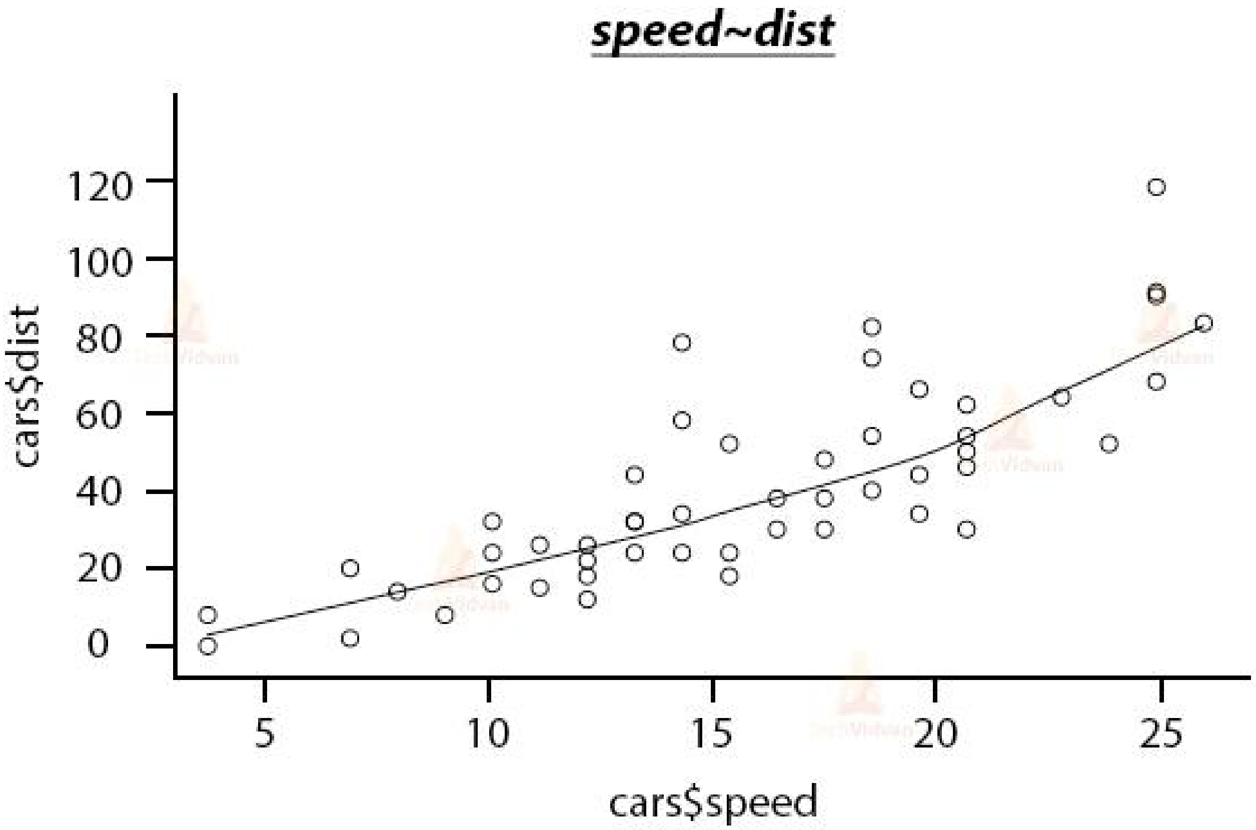


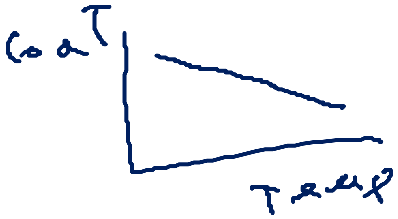
2. Generalized Linear Model (GLM) for Two Values



- narrows down the list of variables to find best fit.
- Generalized Linear Model would narrow down the list of variables, likely suggesting that there is an increase in sales beyond a certain temperature and a decrease or flattening in sales once another temperature is reached.
- advantage of this algorithm is that it trains very quickly.
- clear understanding of how each of the predictors is influencing the outcome.
- requires relatively large data sets and is susceptible to outliers.

2. Generalized Linear Model (GLM) for Two Values



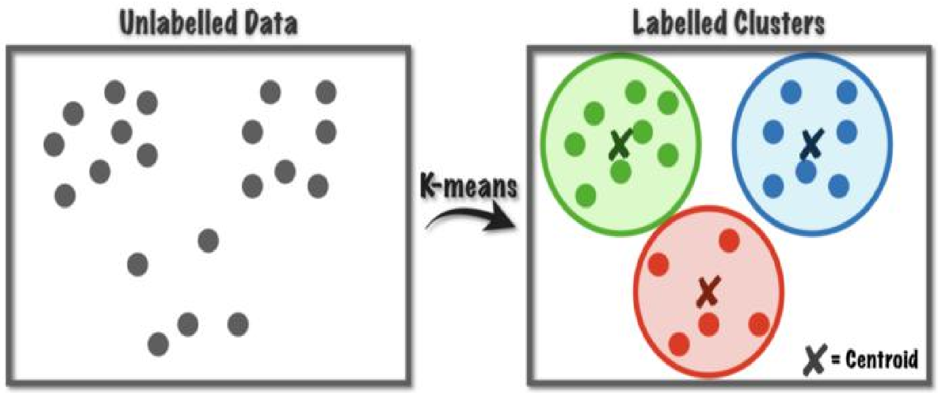


3. Gradient Boosted Model:

- Uses several combined decision trees, but unlike Random Forest, the trees are related. It builds out one tree at a time, thus enabling the next tree to correct flaws in the previous tree. It’s often used in rankings, such as on search engine outputs.
- distinguishing characteristic of the GBM is that it builds its trees one tree at a time. Each new tree helps to correct errors made by the previously trained tree.

4. K-Means:

- Groups data points by similarities and so is often used for the clustering model.
- It can quickly render things like personalized retail offers to individuals within a huge group, such as a million or more customers with a similar liking of lined red wool coats.
- finds clusters such that the observations within each cluster are more similar than the clusters themselves.



5. Prophet:

- used in time-series or forecast models for capacity planning. detects the following trend and seasonality from the data.
- highly flexible and can easily accommodate heuristics and an array of useful assumptions.
- Example: inventory needs, sales quotas and resource allocations.

needs for next month

Benefits and challenges of predictive modelling

- **Benefits:**
- reduce time, effort and costs in forecasting business outcomes.
- environmental factors, competitive intelligence, regulation changes and market conditions can be factored into the mathematical calculation
- **Challenges:**
- much data can skew the calculation and lead to a meaningless or an erroneous outcome.
- massive volumes of data involved in predictive modeling, maintaining security and privacy will also be a challenge.