For Private circulation only

CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Data Geek

By BUSINESS ANALYTICS Specialization
Volume 1/Issue 4/ December 2019

Institute of Management,
CHRIST (Deemed to be University)

CONTENTS

INTRODUCTION TO MACHINE LEARNING ALGORITHMS

“ Google’s self-driving cars and robots get a lot of press, but the company’s real future is in machine learning, the technology that enables computers to get smarter and more personal. ”

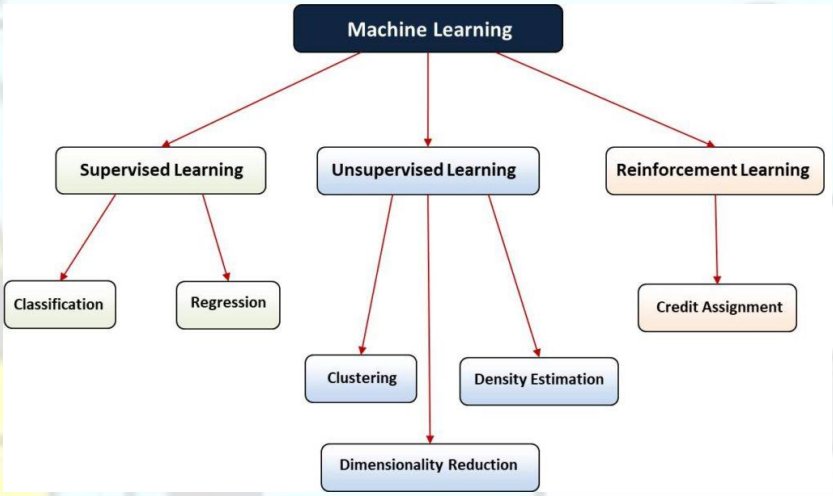
– Eric Schmidt (Google Chairman)

Machine learning being a subfield of artificial intelligence (AI) is also referred to as predictive analytics, or predictive modelling. The term Machine Learning was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence and stated that “it gives computers the ability to learn without being explicitly programmed”. The goal of machine learning is to understand the structure of data and fit that data into models that could be understood and utilized by people. At its most basic form, machine learning uses programmed algorithms that receive and analyse input data to predict output values within an acceptable range. As new data is fed into these algorithms, they learn and optimise their operations to improve performance, developing ‘intelligence’ over time.

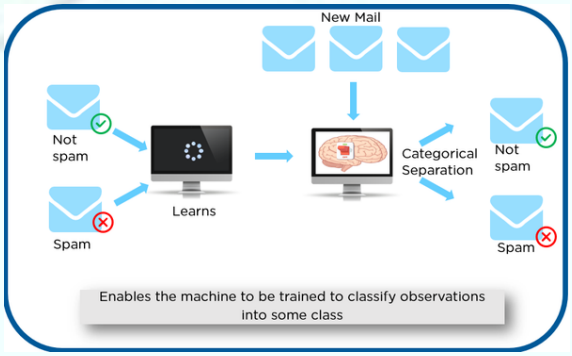
Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

One of the key aspects in machine learning algorithms is selecting the right algorithm for any problem, because there are many algorithms to choose from, understanding their strengths and weaknesses in various business applications is essential. Machine learning algorithms are classified into three major categories, based on how learning is received or how feedback on the learning is given to the system developed.

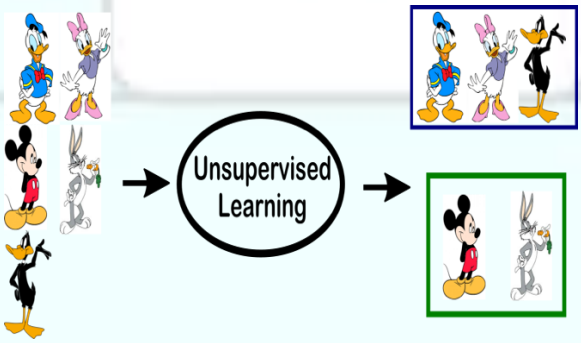
Types of Machine learning algorithms



- Supervised learning- An AI system is presented with data which is labelled, which means that each data tagged with the correct label. The aim is to approximate the mapping function so that the new input data (x) you have can predict the output variables (Y) for that data.
- Unsupervised learning - An AI system is presented with unlabelled, uncategorized data and the system's algorithms act on the data without prior training. The output is dependent upon the coded algorithms.

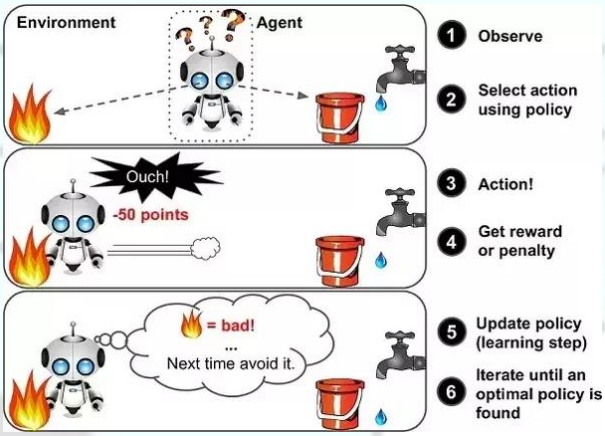


Supervised learning



Unsupervised learning

- Reinforcement learning – AI system learns by interacting with its environment. The agent receives rewards whenever it performs correctly and penalties whenever it performs incorrectly. The agent learns without any intervention from a human thereby maximizes its reward and minimizes its penalty. It is a type of dynamic programming that trains algorithms using a system of reward and punishment.



Reinforcement learning

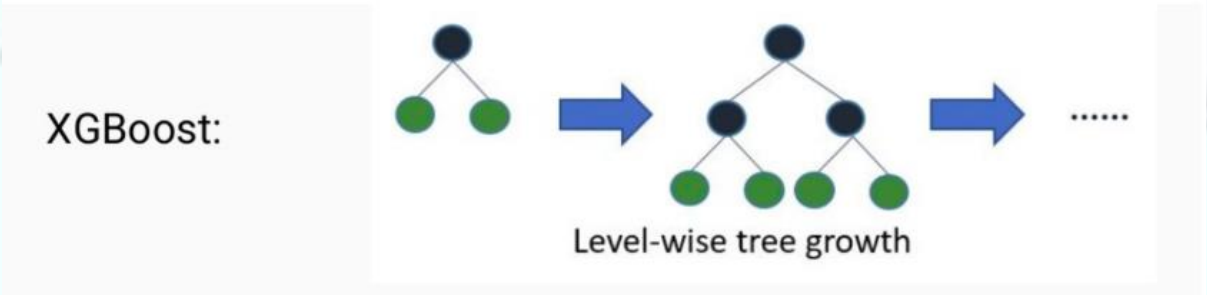
The above machine learning algorithms are classified based on their input type. The application and the usage of an algorithm mostly depend on the problem scenario. Also, Machine Learning is an incredibly powerful field. In the future, it would help to solve most of the pressing problems, as well as open up a whole new world of opportunities in the field of data science.

P.JEYALAKSHMI
1827442

BOOSTING ALGORITHMS IN MACHINE LEARNING

"Mistakes are meant for learning not for repeating."

Boosting Algorithm is a decision- tree based ensemble method for improving the model prediction for any given learning algorithms. In spite of many other machine learning algorithms, boosting algorithms are widely used by the data scientists to boost the accuracy of the model. In simple terms "boosting" refers to converting the weak learners to strong learners in the family of algorithms which include XGBoost, gradient boosting and AdaBoost algorithms. In prediction problems which involve unstructured data, artificial neural network performs the best, while in case of structured data, decision tree based algorithms work the best. However, at present where the data is mostly unstructured, like images or text, boosting algorithms are applied the most in the area of data science. XGBoost uses gradient boosting framework, which has wide variety of tuning parameters. While the methods like bagging, boosting involve minimizing the errors by boosting the performance and predicting based on multiple decision trees, XGBoost algorithm works through parallel processing, that handles the power of multi-core computers across multiple computers as well. It is beneficial when working with large datasets, where it works on GPU, and it is comparatively faster than other ensemble classifiers. For instance, to predict the spam mail using the boosting algorithm. The algorithm initially identifies the weak rule using the different distributions. In this case, if the email has image or has only links it would be classified as spam. These distributions are assigned equal weights, and errors based on prediction are calculated. Based on its higher error rate, it would be considered as a weak rule, and therefore more attention would be provided to it. It becomes an iterative model, until the weak rules are combined to build strong rules, which in turn improves the accuracy of prediction of the model.

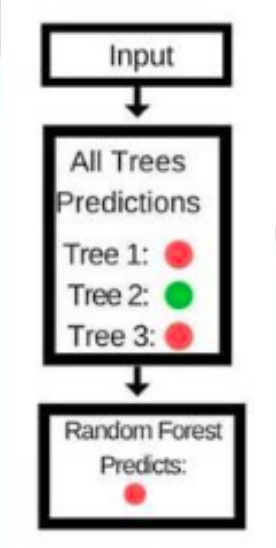


SHWETHA SP
1827048

RANDOM FOREST MODEL

Random forest is one of the widely used Machine learning algorithms. It is used both for Classification (where response variable is Categorical) and Regression (where response variable is Continuous)**.** It uses ensemble mechanism where the variables and the observations are randomly picked, and forms multiple decisions trees. Random Forest addresses problems like Data Preparation, Outliers, Overfitting, and large data.

Before understanding about RFM we need to know how decision tree works. Let us say, we have 9 predictor variables and one Response variable. The decision tree is formed using these 9 predictor variables, and splitted at each node based on Gini index and Information gain values, and finally draws to the response variable class. Here the variables once used at one split cannot be used again for another split. Random Forest is the combination of multiple Decision trees and the average of all the outcomes taken together is calculated, and the final decision is calculated. Here, in each split predictor variables are randomly picked and at each split all predictor variables can be involved that are selected at that instance. The number of randomly picked variables for classification model is by default the square root of the number of Predictor variables, and for regression problem it is the number of predictor variables divided by 3. Based on the Bagging algorithm, multiple samples(training data sets) from the original data with replacement and from each sample 2/3 of the total training data is used for building the model and remaining 1/3 of the training data is used to assess the model performance by calculating the out of bag error(miss-classification rate). Feature Selection is one of the crucial aspects in Random Forest, because when there is multicollinearity between the variables the tree will continuously grow which would increase the complexity of the tree.



Overfitting is the one problem that RFM best handles. Overfitting means when the training data is not able to generalize the patterns, and therefore fails to validate the results. It happens when your training data set is trained too well. So, the Random Forest model works with better accuracy than Decision tree models. While building Random forest, we need to take care of Class imbalance of the response variable. If one class of the response variable has more number of cases, it leads to bias in the model. So we need to under sample the data in such cases. So class imbalance is one of the criteria to be taken care of while implementing random forest algorithm.

MANIKANTH

1827933

UNSUPERVISED LEARNING: CLUSTERING

Unsupervised learning is a technique of Machine Learning where the data to be trained will not have labelled responses. It is usually performed as part of an exploratory data analysis. Unlike Supervised Learning, it is hard to assess the results of unsupervised Learning. The reason for this difference is simple. If we fit a predictive model using a supervised learning technique, then it is possible to check our work by seeing how well our model predicts the response Y on observations not used in fitting the model. However, in unsupervised learning, there is no way to check our work because we don't know the true result.

Techniques for unsupervised learning are of growing importance various fields. The different unsupervised techniques are clustering analysis and factor analysis.

In this article, we elaborate about the most commonly used Unsupervised Learning – Clustering. Clustering is nothing but grouping of observations based on the characteristics they possess. Clustering techniques are used in various domains of businesses –

- Marketing – Market Segmentation: Grouping of customers based on similar behaviour, pattern of buying
- Healthcare- Grouping of patients having similar traits, pharmaceutical drug grouping with similar properties.
- Finance- Grouping of clients based on performance
- Biology- Grouping based on kingdom, phylum, class of animal/plant.
- Media and Entertainment – Building Recommendation Engines.

Clustering Algorithms -

There are many Clustering Algorithms based on the underlying technique of grouping the data points. In this article, three popular clustering algorithms are discussed.

KMeans Clustering

KMeans Clustering is the most commonly used clustering technique. In this technique, the number of clusters to be formed is provided initially and based on that, the clusters are formed by randomly picking centroid. Based on the distance between the data point and centroid, clusters are allocated. In the repeated iterations, centroids are changed by taking mean of all vectors in each group.

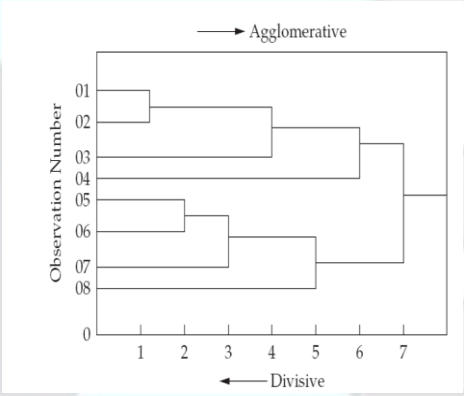
Hierarchical Clustering

There are two types of Hierarchical Clustering: Agglomerative Hierarchical Clustering and Divisive Hierarchical Clustering.

In Agglomerative Clustering, initially each data point is considered as a single cluster. And based on the similarity measure (e.g.: Euclidean Distance, Manhattan Distance) data points with similar features are grouped to form clusters. Then similar groups are merged to form new clusters. This process continues till all the groups are merged to form one single group.

Divisive Hierarchical Clustering is opposite to Agglomerative Clustering where initially all the data points form a single cluster. Based on the similarity measure, groups are divided. The process continuous till the number of clusters are equal to number of observations.

To visualize the Hierarchical Clustering, Dendrogram is used:



DBSCAN Clustering

Density Based Spatial Clustering with Noise is an algorithm which effectively takes care of the outliers. The technique on which DBSCAN works is for every point in cluster, the neighbourhood for a given distance should contain minimum number of points. This algorithm is very effective with a data set having outliers.

The other techniques of Clustering are Mean Shift, Clara, N- Cut, Hybrid, Sting and so on.

NAMRATA MANGALGI

1827648

NAÏVE BAYES ALGORITHM: AN OVERVIEW

What is Naïve Bayes algorithm?

Naïve Bayes is a classification algorithm which works on the principle of conditional probability, as given by the Bayes theorem with fundamental assumptions. The fundamental assumption of this algorithm is that each feature makes an **independent** and **equal** contribution to the outcome.

Bayes' Theorem calculates the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is mathematically denoted as:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Here A and B are two events and,

- P(A|B) : the conditional probability that event A occurs , given that B has occurred. This is also known as the posterior probability.
- P(A) and P(B) : probability of A and B without regard of one other.
- P(B|A) : the conditional probability that event B occurs , given that A has occurred.

Naïve Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, it is also known to outperform even highly sophisticated classification methods.

Understanding Naïve Bayes classification

Based on the Bayes theorem, the Naïve Bayes Classifier gives the conditional probability of an event A given event B has occurred. Let us understand the same using an example:

- Problem statement: To predict whether a person will purchase a product on a specific combination of day, discount, and free delivery using a Naïve Bayes classifier.

- Solution: With the help of the variables under day (weekday, weekend and holiday), for any given day we can check if there are any discounts and free delivery. Based on probabilities calculated with parameters we can classify customers as 'buyers' and 'not buyers'.

Important applications of Naïve Bayes algorithm

1. **Text classification :** Naïve Bayes classifiers are mostly used in text classification (due to their better results in multi-class problems and independence rule) and have a higher success rate as compared to other algorithms.
2. **Spam Filtering and Sentiment Analysis:** . With its higher success rate it is also used in spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments).
3. **Medical Diagnosis :** Naïve Bayes classifiers also assist doctors to diagnose patients by using the information that the classifier provides. Naïve Bayes can also indicate if a patient is at high risk for certain diseases and conditions, such as heart disease, cancer, and other ailments.
4. **Recommendation System:** Naïve Bayes Classifier along with Collaborative Filtering algorithms makes a Recommendation System which uses techniques like machine learning and data mining in order to filter a particular information and predict whether a user's preferences
5. **Image classification** : With its high efficiency in handling binary (cat, not cat) and multi-class classification (cat, dog, mouse), Naïve Bayes is widely used for image classification.

P.JEYALAKSHMI
1827442

PROJECT ALBUM

This section is an attempt to delineate a few interesting projects taken up by the students of MBA Business Analytics of CHRIST Institute of Management. This spans around various themes such as Customer acquisition to Credit card fraud to health analytics to name a few. The insights generated from data indeed is a proof of the amazing skill sets that our students have earned from the courses as part of the BA curriculum.

Predictive Analytics: A way to curb Credit Card Default in banks

Credit card defaulting by the customers is one of the biggest challenges banks face, mainly in the consumer banking segment. Every credit card issuer wants to predict the likelihood of default by customers. For the purpose, banks need to know the factors that lead to and determine the likelihood of defaulting. Based on the key drivers and prediction, the credit card issuer can decide who to be issued a credit card and the credit limit to be provided. The prediction also helps the issuer to recognise their current and future potential customers, and to strategize credit card offers accordingly.

My Work covers the credit default prediction of Taiwanese bank based on the data of 30000 credit card customers. The classification of whether a customer is defaulter or not is done based on 24 parameters like demographics, credit data, and history of payment and bill statements of credit card customers for previous 6months etc. During the process of model building, it was found that the factors such as Gender, Education, Marital Status, Repayment status of the previous months contribute in the better model prediction.

Algorithms like Logistic Regression, Decision tree, Random Forest were employed to build the prediction model. Of all, the logistic regression is found to deliver better accuracy of 82%. In this case, Binary Regression models will be able to classify better and predict whether a customer will be defaulter or not.

MONICA CHOWDARY
1827042

Visualization of employee attrition at IBM

Experts today loudly proclaim the potential of workforce analytics to improve workforce productivity, enhance employee experience and wellbeing, and increase the impact of HR. But if these possibilities are to be fulfilled, organizations must take certain steps to ensure the right foundations are in place. In companies like IBM, the employee attrition has been noticed and hence recorded. The dataset is created by the IBM data scientists for studying. It has 34 factors like age of the employee, his/ her marital status, the department etc. Main objective is to visualize the data, analyse them using suitable charts and identify the main reasons leading to employee attrition. The given dataset is analysed to derive dimensions and measures to solve the problem. Dimensions are Department, Marital Status, Educational Field and Job Involvement. IBM Cognos is very useful tool for analysing huge amount of data. It helped the HR team to take appropriate decisions based on understanding the data in different dimensions. Multi – dimensional perspective of data helps in identifying the exact reasons for problem and even helps to find solutions. The solutions were that the company should try to reduce overtime in R&D and Sales department in order to retain employees. The employees should feel connected with the company increasing their job involvement. Lastly, the married male employees should be given incentives according to their positions in the company to prevent them from switching jobs.

MIHIKA CHATTOPADHYAY
1827040

Leveraging Predictive analytics for customer Acquisition/retention in Banking Sector

Customer acquisition/retention is a compelling challenge faced by the consumer banking sector. Banking/Financing institutions are attracting customers with generous and rewarding promotional offers .As banks aggressively spend on customer acquisition; they must develop strategies to create value out of their expenditures. They must focus on acquiring right customers and build long-term relationships with them. Banks primarily treat customers under two categories: liability customers and asset customers. Banks want to convert the liability customers to asset customers in order to improve profitability.

Retail marketing department has to devise campaigns and do promotional campaigns only for the potential customers which will increase the success ratio with minimal budget.

My focus of the work was on one of the use cases of banking sector: to identify the likelihood of the customer to purchase personal loan for Thera bank. The data available is the list of liability customers up to 5000 customers of the bank and the variables defining the credit card spending, mortgage with the bank, years of experience with respect to customer's occupation, size of the family, customer's possession of Cash Deposit account, Income of the customers, Education level and age etc. supervised machine learn techniques like logistic regression, random forest and decision trees are used to uncover various insights from the data.

All the models exercised under this use case are accurate, where in random forest has reached the maximum accuracy at 1. Distinguishing power among these models is higher with logistic regression model with an AUC score of 0.9633.Insights uncovered from these models explains that the factors like income and possessing cash deposit account are more likely to purchase the personal loan from the bank. Decision trees also throw light on various factors leading to the customers purchasing a personal loan which will help marketing team to devise campaigns in a focussed approach. These insights help build strategies to cross sell or upsell various products.

MSRK PATANJALI
1827313

Business Intelligence and Visualization of Sales Data

The article focuses on the need of Business Intelligence and Visualization in the area of selling of sports accessories in a company. For companies which are widespread across the world, it's difficult to manage operations and trace the regions which are not performing well. It is a cost for the company to manage the operations in the stores which aren't making any profits. So, the approach to this problem is the use of Business Intelligence tools and techniques effectively.

By making use of data provided by the company we have identified the dimensions and measures to draw a star schema and relate the variables across various tables. In order to have a better visualization across regions, we have merged the countries and came up with a new column which helps to focus on a much broader level of sales of equipment. IBM Cognos was used for visualization of sales across regions. In order to do this, analysis of the trend in the sales across various regions is required. So, various charts like bar chart, line graphs and pie

charts were used to represent and analyse the data. Line chart basically gives the trend and helps to analyse the fluctuations in the sales. Here the sales is compared with respect to profits region wise, thereby identifying the regions making zero profits.

The profits made from various distributers and the sales made by them were also identified. There were some distributers whose growth was dropping since the past few years. Also, some of the products were on high demand, hence the demand could be forecasted, and the inventory level be maintained accordingly.

So, using the BI tools helped to give recommendations by identifying the trend in profits and sales of the equipment, and the steps to be taken to help them reduce the operation costs.

KAMAL SUNIL
1827111

Predictive Analytics: Dengue prevalence, by administrative region

Machine learning algorithms are widely used in various fields today. Healthcare is one of the industries where the implementation of these techniques is found to be very high, in order to diagnose a disease based on the factors that influence or cause a disease. With respect to this, a prediction model is built with an objective of predicting the presence of dengue for a particular region by taking into consideration some of the geographical factors.

Data are recorded for each of the 2000 administrative regions to check if Dengue or not was recorded at any time between 1961 and 1990 or not. The classification whether a region has recorded dengue or not is made based on the 13 variables, which mainly include the factors such as humidity, temperature, density of forest, latitude and longitude. These variables played a major role in the spreading of dengue, and contributed to a better model prediction.

Models were built based on the classification techniques such as Logistic Regression, Decision tree and Random Forest. Out of these models, Logistic regression model and Random Forest exhibited the accuracy of almost 95% whereas the Decision tree model exhibited the accuracy of 89%. Hence, the Logistic regression and Random forest models can be employed to predict whether a region would be affected by Dengue or not given the factors mentioned above are found to be present in a particular region.

CHANDAN A J
1827007

Visualization on video games using BI tools

The video game industry has evolved tremendously over the last 30 years. The games designed initially were mechanically and graphically very simple. Due to the technological advancements over the years have led to the development of virtual and 3D games. As the market has developed drastically, not every game sees a breakthrough. My project aims at analysing the Sales Video Games Industry that would help a new gaming company to design their strategies to create a niche for itself in the market. The analysis includes the Data Visualization Techniques to identify the customer behaviour and pattern of the video game purchases. With the use of Tableau and IBM Cognos, I was able to visualize patterns at the granular level that helped me to understand the business better. The Overall Sales has seen a decreasing trend, denoting that the advanced technology is influencing customers towards Virtual and Augmented Reality games. Recommendations to develop Action and Sports games, into more adventurous ones with application of 3D Augmented Reality techniques on the PS2 platform targeting all the age groups at the global level could be focused by the company to sustain and compete with the other players in the market. Tableau and IBM Cognos provided good learning experiences with better understanding of the figures to infer about the sales, comparing it with other external factors as well.

SHWETHA SP
1827048

Analysis of the performance of car sales representatives through visualisation

Business Intelligence tools are used to transform the data into meaningful visuals, reports, or dashboards that provide context for analysts to make business decisions. This is a very important part of any business as business decisions needs to be taken after analysing the data which the business has in hand.

Subaru an automobile manufacturing division of Japanese transportation, a conglomerate Subaru Corporation. Toyota owns around 16% of Fuji Heavy Industries which is the parent company of Subaru. Toyota and Subaru are the two makes which are being sold by national dealers with few sales representatives in USA. The main objective for my work being to analyse the performance of sales representatives of a dealer selling Toyota and Subaru vehicles.

For managers and corporate executives to track and analyse factors deemed crucial to the success of an organization or a company, KPI's are to be mentioned. For selling a vehicle, sales representatives have to be very convincing and a lot of nudging might be needed for a deal to be wrapped, so rewarding of the sales reps is necessary for them to perform better every time. By using the IBM Cognos tool, managers can analyse how each sales representative is performing and based on that he/she can decide on how to reward them.

By analysing the data using IBM Cognos, it was seen that total sales profit of Toyota is approximately double of Subaru. It was seen that few sales representatives were more focused towards selling Toyota cars as more profits was linked to Toyota cars. Few sales representatives made high profit for company by selling fewer cars but which was of higher value, this should be noted by the manager and accordingly they could be rewarded. Similarly few cars are of lesser value compared to the cars which are of higher value, the sales representatives who sell those should also needs to be recognized and rewarded. Decisions on which sales representative gets the more rewards lies with the manager, by analysing the data he/she can make a well informed decision. By using IBM Cognos the analysis could be drilled down to get more granularity. It's a powerful tool to create dashboards and reports which in fact helps in taking informed decisions.

DEEPTHIPRIYA R PILLAI
1827034

A DAY WITH A DATA SCIENTIST

Mr. VENKAT RAMAN, TRUE INFLUENCE

Mr. Venkat Raman is a Data science professional, being a data scientist he has developed several Data Science products using machine learning / NLP techniques and several proprietary statistical techniques. Currently he is working as Principal Data Scientist at True Influence, where he is in charge of developing Data Science Solutions for the Insight BASE Platform.
True Influence is a data-driven technology company that connects the customer. They expertly leverage data, technology and content to drive high-impact marketing campaigns and share detailed data insights to help win new business. True Influence drives leads and generates revenue across multiple industries, promoting brands and products from many of the most successful US companies. While they primarily serve large and sophisticated marketing departments such as those at IBM, Microsoft, Google, Marketo, Oracle/Eloqua, Symantec and more.

1. Can you tell us about your designation and about the company?

I work at True Influence as a Data Scientist. True Influence is a US based Marketing tech company. The company was founded in 2008. It has offices in India, USA and UK.

2. Can you elaborate about the company's primary focus areas? (in terms of sectors and technologies)

True Influence specialises in B2B marketing automation software category, lead generation and Online Intent monitoring. As a Data driven company, True Influence deals with the 'Big Data'. To handle such huge Big Data and extract meaningful insights, the company uses

various cutting-edge technologies like AWS cloud, Spark, EMR, Hadoop, Machine Learning, NLP and other proprietary technologies developed inhouse.

3. For Analytics professionals which plays a major role either domain knowledge or expertise in tools and techniques?

It is a very good question. Which plays a major role – 'domain knowledge or tools expertise' really depends on at what stage of career the analytics professional is. Domain knowledge comes from years of experience whereas mastering a tool or technique takes less time (relatively speaking). As one climbs the corporate ladder, one would realize that tools are a means to end and not an end itself. This doesn't belittle the importance of tools, it has its utility but the knowledge of what to do with the tool is paramount. The latter is where domain knowledge has an edge over expertise in tools/techniques.

It is important for young inexperienced analytics professional to be extremely good at tools/techniques. This is what will open up more opportunities early in the career. As one gains more experience, with it will come domain experience.

Once a person has substantial years of experience under the belt, domain knowledge will matter more as one would be expected to mentor or guide the junior professionals on what to do with tools and when to apply which tool.

4. Sir, you have won "outstanding performance" award for a work on devising ranking criteria through data analytics from online questionnaire designed for CIO 100 2014 event, while working as technology marketer in IDG, did this milestone lead to a career transition for you from marketer to data scientist.

I would say it was one of the significant moment which made me believe I could become a data scientist. For the first time I had applied analytics techniques to a business problem and luckily it worked out just fine.

5. What was your inspiration for being able to achieve this milestone? What were the challenges you faced during this transition?

The inspiration was just to prove my mettle to my company (IDG) which had reposed faith in my skills.

There were many challenges.

- How to structure the questionnaire in a logical manner
- How to collect the answers from over 500 companies, clean and standardize the data
- How to code the categorical answers
- What appropriate ranking logic to devise to rank the top 100 companies and how to sub rank these 100 companies under their respective verticals.

All these are typical problems which are faced by any data scientist but as a green horn, way back in 2014 these problems were really new to me. I guess, as a single person working on such a huge problem and ultimately coming up with goods gave me an immense sense of satisfaction and confidence.

6. 5 years down the line where do you think will be the future of analytics?

One thing that any seasoned data scientist would tell you is the fact that, prediction is a very tough business. One can very rarely predict correctly. 5 years into the future is too long a time to predict. But I could perhaps say where the Data science as an Industry is directionally headed. Gartner releases Hype cycle chart every year. As with any new Industry, there is "peak of inflated expectation" followed by "disillusionment" and then "enlightenment" and productivity.

In my opinion, from 2010 – till present, we went through the states of "inflated expectation" as well as certain "disillusionment". I think in next 5 years, Data science will be more readily accepted and implemented across various industries, perhaps a combination of "enlightenment as well as realising productivity".

7. What are the major challenges faced by analytics professionals?

I would like to bucket the challenges faced by analytics professionals into two categories

- Data Problem
- Stakeholder Management Problem

Data problem

Under Data problem, one would typically find issues like messy data, huge volume of data, what algorithm/technique to be used, lack of adequate data processing resources (servers, cloud), Database issues etc.

Stakeholder management problem

One of the major challenge for a Data Scientist is that - How to make the non-data science background stakeholders understand that your algorithm/ Data science solution is remarkable.

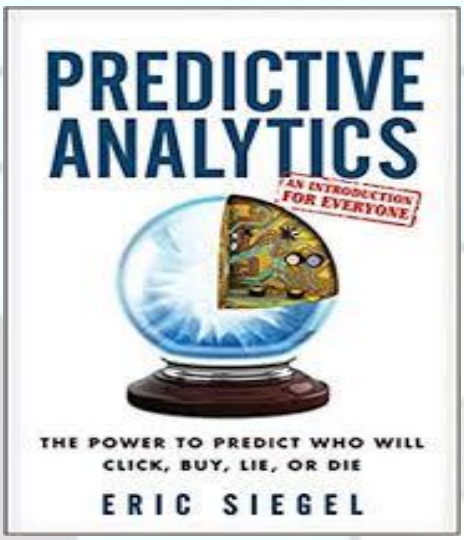
Everything from Job promotions, Hikes, perks, options in the company, funding for your project, grant to apply for patents etc depend on whether the stakeholders really understand the value you bring to the table.

The other issue is with expectation management, most often stakeholders start viewing data science as a 'silver bullet'. This unrealistic expectation leads stakeholders to the belief that they are not getting the ROI.

Most of the time Data problems are solvable and they are not a career threatening challenge. Whereas Stakeholder management problem is a really tough challenge.

PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE

(By Eric Siegel)



Eric Siegel is the founder of Predictive Analytics World and executive editor of The Predictive Analytics Times. He is a former Columbia University professor, and is a renowned speaker, educator, and leader in the field. His book reveals through a plethora of cases how predictive analytics works, and how it affects everyone every day.

As mentioned throughout the book, predictive analytics reinvents industries and it claims to run the world, more so in today's world where every industry is inundated with data. A prime example is the life insurance industry whose every decision is based on prediction.

Eric Siegel uses a wide variety of applications to explain what predictive analytics is, and the way it is used in industries to get the optimum output. Without delving into the mathematical and the technical aspect of the area, he described the practical application of prediction in various businesses. This makes the book an easy read for people across domains. Instances include such as Hewlett-Packard forecasting which of the employees are likely to quit the job; Chase Bank forecasting which mortgage customers are about to refinance and switch to another firm; Brigham Young University Hospital predicting premature births et al. This would be the

go to book for those who want to understand predictive analytics in general terms. The emphasis throughout the book has been to explain the subjects using the least jargons possible. Siegel claims that, “a little prediction goes a long way.” This has been an impetus for him to be associated with the field, and make people aware of the power of prediction.

He devotes a chapter on IBM Watson, wherein he described the entire episode of how the computer beat human champions on the game show Jeopardy. Watson was designed to come with possible answers to a probable question to be asked in any game-show, and then “predict” which one was the best one. The prediction in this case became its final answer leading to the defeat of the two former Jeopardy champions.

The book also has a chapter on the concepts of machine learning, focusing mainly decision trees. It also has some tables that list various surprising insights of predictive analytics. One such insight mentioned is, Orbitz likely shows expensive options to Mac users, because it learned that Mac users spend up to 30% more money than Windows users.

The author also describes certain important moral and ethical aspects that rises due to the increasing ability to predict individual behaviours. He ends the book by discussing the future developments in the field of predictive analytics and how would it be by the year 2020, and is affirmative of the fact that it would influence the lives of billions of people.

GAUTAM DEKA
1827208

Quiz Corner

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

a. Decision Tree
b. Regression
c. Classification
d. Random Forest

2. To find the minimum or the maximum of a function, we set the gradient to zero because:

a. The value of the gradient at extrema of a function is always zero
b. Depends on the type of problem
c. Both A and B
d. None of the above

3. The most widely used metrics and tools to assess a classification model are:

a. Confusion matrix
b. Cost-sensitive accuracy
c. Area under the ROC curve
d. All of the above

4. Which of the following is a good test dataset characteristic?

a. Large enough to yield meaningful results
b. Is representative of the dataset as a whole
c. Both A and B
d. None of the above

5. Which of the following is a disadvantage of decision trees?

a. Factor analysis
b. Decision trees are robust to outliers
c. Decision trees are prone to be overfit
d. None of the above

6. How do you handle missing or corrupted data in a dataset?

a. Drop missing rows or columns
b. Replace missing values with mean/median/mode
c. Assign a unique category to missing values
d. All of the above

7. What is the purpose of performing cross-validation?

a. To assess the predictive performance of the models
b. To judge how the trained model performs outside the sample on test data
c. Both A and B

8. Why is second order differencing in time series needed?

a. To remove stationarity
b. To find the maxima or minima at the local point
c. Both A and B
d. None of the above

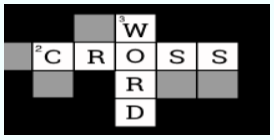
9. Which of the following is true about Naive Bayes ?

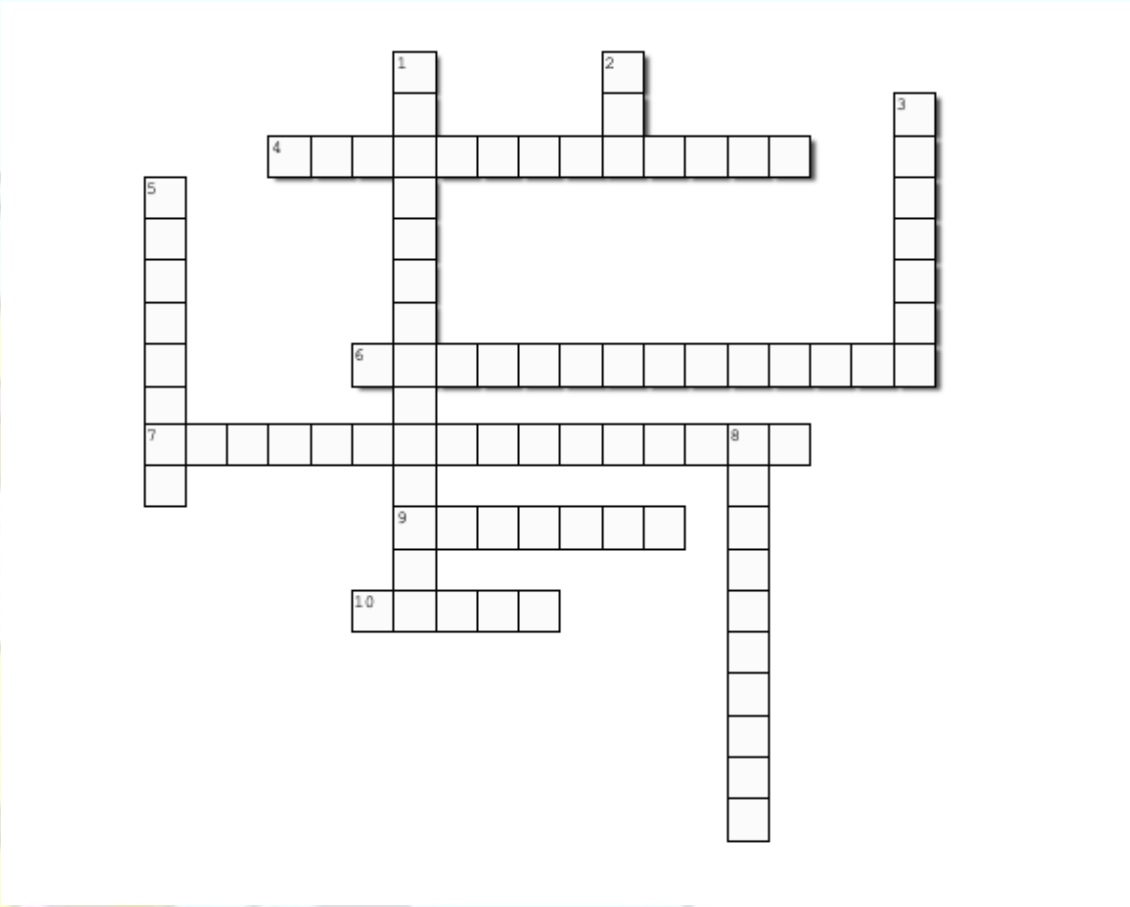
a. Assumes that all the features in a dataset are equally important
b. Assumes that all the features in a dataset are independent
c. Both A and B
d. None of the above options

10. Which of the following techniques can be used for normalization in text mining?

a. Stemming
b. Lemmatization
c. Stop Word Removal
d. Both A and B

Crossword





ACROSS:

4. Process of rescaling any data
6. Points which are actually false but are incorrectly predicted as true
7. A table that is often used to describe the performance of a classification model
9. Technique where final predictions are determined by combining the combinations of multiple models
10. An open source, high level neural network library, written in python

DOWN:

1. Another name for Boolean variable
2. Initial step to summarize main characteristics of a dataset
3. A measure of effectiveness of classification
5. A metric by which one can examine how good id the machine learning model
8. Technique used for handling missing values in the data

ANSWERS

ACROSS-

4. Normalization
6. False positive
7. Confusion matrix
9. Bagging
10. Keras

DOWN-

1. Dummy variable
2. EDA
3. F score
5.Accuracy
8.Imputation

DATAGEEK CREW

STUDENT COORDINATORS

JEYALAKSHMI

GAUTAM DEKA

JURNOs

CHANDAN AJ

MANIKANTH

MONICA CHOWDARY

NAMRATA MANGALGI

MSRK PATANJALI

SHWETHA SP

KAMAL SUNIL

DEEPTHIPRIYA R PILLAI

MIHIKA CHATTOPADHYAY

INTERVIEW

ARCHANA SCARIA

HARITHA T

DESIGN AND CREATIVITY

SHREYA MISHRA