greatlearning
Learning for Life

DSBA Project List

A comprehensive collection of the top projects done by our learners from various domains.

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

GREAT LAKES
EXECUTIVE LEARNING

1.1 Optimization of Equity Portfolio

Portfolio optimization is a formal approach to making investment decisions across a collection of financial instruments or assets. The portfolio optimization problem lies in minimizing the risk of an investment for a desired level of expected return. It is estimating asset return and total return moments from price or return data. The idea behind an "optimal portfolio" comes from the modern portfolio theory wherein this theory assumes that investors focus their efforts on minimizing risk while also striving to attain the highest possible return. The business problem here is to provide investment advisory to a client basis his risk profile and his expectation of return. Advisory is limited to equity asset class only, that too, for Indian Market. In this project, data from NSE 500, moneycontrol.com and valuresearchonline.com has been used.
As part of the project, we have done selection of 50 stocks from NSE 500 and have assigned relative ranking to them based on the financial and fundamental data. Further, we have used non-linear optimisation techniques and used 10 of these 50 stocks to create a stock portfolio for optimising returns within a required maximum loss. We have done prediction modelling for stock price which would be able to predict the stock price for t+3 days and t+10 days fairly accurately. We have used the techniques of linear, Ridge and Lasso regression and non-linear programming in the project. The prediction model for stock price has been done using the programming language 'R' and the domain of the project is Finance Risk Analytics.

***Keywords**: Portfolio Optimization, Non-linear programming, Finance Risk Analytics*

1.2 Analytics for a Retail Fashion Brand

The purpose of this capstone is to assess the practical usage of analytical techniques in solving a real-world business problem. This study has analysed various facets of a retail fashion brand with an objective to increase its presence and profitability in a highly competitive fashion industry. The brand has a wide range of progressive seamless wear to maximize flexibility, mobility and performance while enhancing comfort and confidence. The Management has restrained us from disclosing/ using the name of the brands and the Company in any kinds of reports/ presentations whatsoever. The vision of the brand is to symbolize the multiple facets of women's life which she seamlessly handles in different roles of her life. Aim of each product of the brand is to truly bring alive the philosophy of "Born Free" with fabric and design that makes it a fashion statement and even a personal statement. The said brand is a new entrant in this domain and in a very adolescent stage where it needs a proper and a guided approach to be amongst the best and leading women's fashion brand. The fashion industry is very competitive and to make a presence, a smart approach is required along with the hard work. To compete with the humungous fashion brands available today, the Company has decided to approach analytical methods to get the best out of the business. Analytics can help them unearth the insights that would have otherwise remained hidden. This project will help them make informed decisions that will ultimately shape the growth of the business and it will help us to understand efficacy of using analytics in real time data. During the project, three tools were pre-dominantly used viz. Microsoft Excel, Tableau and R. MS Excel is mainly used for cleaning the data and understanding different variables. Tableau is widely used for Exploratory Data Analysis and gathering visual insights. R is primarily used for statistical modelling.

***Keywords**: Retail Fashion Brand, Flexibility, Mobility, Real Time Data, Statistical Modelling*

1.3 Sales Analytics on certified pre-owned cars in U.S

Cars are being sold more than ever. Countries adopt the lease culture instead of buying a new car due to affordability. Therefore, the rise of used cars sales is exponentially increasing. Car sellers sometimes take advantage of this scenario by listing unrealistic prices owing to the demand. Therefore, arises a need for a model that can assign a price for a vehicle by evaluating its features taking the prices of other cars into

consideration. In this project, we have use supervised learning method namely Random Forest and Regression technique to predict the prices of used cars. The
model has been chosen after careful exploratory data analysis to determine the impact of each feature on price. A Random Forest with 500 Decision Trees were created to train the data. From experimental results, the training accuracy was found out to be 74 %, The model can predict the price of cars accurately by choosing the most correlated features.

***Keywords**: Exponential Increase, Supervised Learning, Random Forest, Regression*

1.4 Prediction of policy buyers and their segmentation

An insurance policy lapse is a major concern for most of the insurance companies as it generally occurs within the first policy year and prevents insurers from recovering initial expenses of the policy acquisition. Insurance companies have started focusing on retaining the customers, reducing the lapse rates and hence categorizing them and creating a specific set of customers depending on the various factors as per the policy outlined by each insurance company. The focus on any cross sell would be only on these specific customer profiles due a heavy cost of acquisition. A customer's life cycle in the insurance business domain depends on the factors like age, gender, marital status, income, insurance, density of the place etc., Analytics can be used to assess a customer's total life time value. Here we would need to achieve the following objectives:

a) Predict if the customer would be willing to buy or not buy the policy.
b) Segment these customers with into proper class like Platinum, Gold, Silver.

By doing the above, we hope that this will increase their product sell and in-turn make their marketing efforts impactful and effective. In any business, the return on investment is a key metric for assessing a product profitability.

***Keywords**: Policy Lapse, Customer Retention, Product Profitability*

1.5 Prediction and the rule of law- justice simplified

This project focuses on a class of "Machine Learning" techniques and their potential impact in the domain of Legal Analytics. This work in no way to be considered as a suggestion that all or most of the tasks routinely performed by attorneys are automatable. The tools and techniques used to substantiate our findings include working on R Studio and using Data Mining, Predictive Modelling and Machine Learning Techniques limited to use of Decision Tree, Logistic Regression, Multinomial Logistic Regression and K Nearest Neighbours (KNN). We have limited our study to last 20 years' data and 2 decision variables: Case Disposition and Decision Type. The findings of this study gives the insight that KNN is one of the judicious technique for the treatment of missing values, in case of multi-dimensional data. We found out that Decision Tree algorithm accuracy, precision and recall is better than Logistic Regression and Multinomial Regression for both the decision variables. Cross Validation technique is also used to verify the consistency of the model performance. At the end we have been successful to find the core features for a base model in the Legal Analytics domain, which can be extended further for a deeper understanding of the legal domain and building a robust and sustainable model.

***Keywords**: Machine Learning, Legal Analytics, Data Minding, Predictive Modelling, KNN, Cross Validation*

1. 6 Market Penetration – An Analytical Approach

The project is about a body wash product 'ThinkSkin' which is a private label for a leading retail organization. Brand currently enjoys 16% of market share. The organization wants to set a penetration guidance of 25% in the industry, based on the past purchase patterns and the type of customer Walk-in. The objective of this project is to develop an analytic framework to generate actionable business recommendations to enable growth of the category and become market leader in the near future.

***Keywords**: Retail Organisation, Market Share, Penetration Guidance, Analytic Framework*

1.7 Portfolio Risk Assessment and Analysis

In this report we try to understand the relation between Stock Risks 'β' in a portfolio, investors Risk tolerance level 'λ' & how investors have invested accordingly. We start the problem by understanding the risks involved in the stock held by the investor, followed by their risk tolerance assessment and finally by understanding the relationship between the two metrics. The Markowitz Efficient Frontier inspired us in understanding that, by simply assessing the stock or market risk is not enough to have better understanding of the customers behavioural tendency with fluctuations in market. We thereby, illustrate the underlying relationship in a tabular format to get a good view of the problem.

***Keywords**: Stock Risks, Risk Tolerance, Markowitz Efficient Frontier*

1.8 Marketing Analytics for Bajaj Allianz Life Insurance

Bajaj Allianz Life Insurance Ltd undertakes regular email campaigns to increase brand awareness, sale of insurance policies and to create unique customer touch points. The core objective of the study is to use various analytics techniques to deep dive into the responses and aim to optimize the hit rate of these campaigns. Techniques used here are Exploratory data analysis techniques, Classification algorithms including Decision Trees and Logistic, Machine learning techniques like Clustering, Neural Network, CART, SMOTE, Random Forest, Gradient Boosting and Under sampling and oversampling.

Keywords: *Marketing analytics, Customer analytics, Random Forest, CART, Neural Networks, Gradient Boosting and Clustering*

1.9 Predicting Fate of Indian Start-Ups

India has the third highest number of start-up incubators and accelerators in the world, after China and the US. The number of start-ups and incubators grew by 40 percent in 2016-17. Of these, 30 academic incubators were established under the government's 'Start-up India, Stand-up India' initiative. The idea was simple-simplifying compliance and roll out an action plan for giving a fillip to innovation and entrepreneurship in the country. In this direction, the government announced measures such as compliance based on self-certification, a mobile app to provide on-the-go accessibility for registering start-ups with relevant agencies of the Government, creation of a Start-up India hub, legal support and fast-tracking patent examination at lower costs, tax exemption for start-ups, providing funding support through a Fund of Funds, and many other initiatives. Government support, technology enhancement and development in metro cities has boosted the start-up ecosystem. This has attracted numerous numbers of investors, both national and globally. Therefore, a large amount of money is poured into the start-up ecosystem. Start-up companies' expenses tend to exceed their revenues as they work on developing, testing & marketing their idea and need funds to expand their business at different stages of development. This project aims at analysing Indian start-ups from funding, failure and profitability perspective and to predict the outcome of start-ups in India.

***Keywords**: Start-ups, Incubators, Legal Support, Fast-tracking, Tax Exemption, Funding, Profitability*

2.1 Analysis of Factors Affecting the AQI of Delhi

New Delhi is among the most polluted cities in the world today. Air pollution is responsible for many health problems in the urban areas. Of late, the air pollution status in Delhi has undergone many changes in terms of the levels of pollutants and the control measures taken to reduce them. The situation worsens during the onset of winters every year from October onwards. There are various causes contributing to the current state of air pollution in Delhi.
There were conflicting reports on media on the actual cause of air pollution in New Delhi. Through this study we hope to develop some insights that can help organizations (State/Central Pollution Control Boards and NGOs) to advocate more stringent policy frame work to control air pollution. There are multiple factors which affects the air quality, such as: stubble burning in NCR, construction activities, vehicular movement, firecrackers, industrial pollution, diesel gensets, etc. But, the data for all of these is not available readily. We could collate the metrological data from CPCB & an estimate of vehicular traffic through google maps, & thus decided to restrict our study only to these two sources of data.
An estimate of pollution caused by vehicular movement can be studied if the data on vehicles running on road at any given point of time is available. However, Google Maps provide only data in terms of time it takes to move from point A to point B. Since, the speed of vehicles on road majorly depends on the density of vehicles, we decided to take speed of vehicles as an estimate of vehicular density. The same was calculated using the time & distance we got from the google maps.

***Keywords**: Air Pollution, Metrological Data, Vehicular Movement*

2.2 Predicting Graduate Engineers Employability

Graduate employability is an increasingly major concern for academic institutions and assessing student employability provides a way of linking student skills and employer business requirements. In the last four years, there is no significant improvement in employability of engineers. Recent study by Aspiring Minds NRE Report shows that only 17.91% of engineers were employable for the software services sector, 3.67% for software products and 40.57% for a non-functional role such as Business Process Outsourcing. Student's employability is a major concern for the institutions and predicting their employability beforehand can help in taking timely actions in order to increase institutional placement ratio. To know weakness before appearing for interview of any company can help students to work in areas that they need to improve in order to best match the skillset required by company. Enhancing student assessment methods for employability can improve their understanding about companies' in order to get suitable company for them. Data mining and predictive modelling technique such as classification and regression is best suited for predicting the employability of students. The application of data mining in student employability is to search for significant relationships such as patterns, association and changes among variables in datasets. It provides classification methods to predict the level of employability for students.

***Keywords**: Graduate Employability, Software Services, Business Process Outsourcing, Data Mining, Predictive Modelling*

2.3 Study of Family Planning Methods by Indian Women

SHOPS Plus is promoting the adoption of safer sexual practices, including delaying sexual debut and using modern contraception to avoid unintended pregnancies. Our project is a subset of this project and aims

to determine patterns in the selection of family planning methods by Indian women using data analytics techniques.
The groups obtained by analysing the data will reflect the choice of contraceptive given a woman's socio-economic and demographic indicators. These group definitions can be used to promote relevant family planning products and schemes by the government. We will also develop a predictive model using the information obtained above. However, in absence of any new information to test the model, it may only serve an academic purpose at best.

***Keywords**: Safer Sexual Practices, Data Analytics, Socio-Economic, Demographic, Indicators*

2.4 Credit Risk Prediction, Customer Segmentation and Roll Rate Analysis

Credit risk profiling is very important for banks and other lending institutions. Profiling risky segments can reveal useful information for credit risk management. They decide who is credit worthy and who is not based on the individuals'/companies demographic information, credit history with the bank (if available) and the information available at Credit Bureaus. Credit providers often collect a vast amount of information on credit users. Information on credit users (or borrowers) often consists of dozens or even hundreds of variables, involving both categorical and numerical data with noisy information.

This is a typical classification project where the company seeks to identify which clients are creditworthy. The project also seeks to provide the following solutions to the company:

- Segment customers such that the company can develop marketing strategies to target specific segments which will increase the potency of the marketing strategy
- Compute roll rates to predict credit losses based on delinquency

2.5 Automatic Summarization of News Articles

Text summarization can be done in two ways: Extractive text summarization, where the algorithm picks important words from the article to generate a summary of the article, and
Abstractive text summarization where algorithm generates a summary of the articles with its own words just like a human generated summary.
Automatic summarization of news articles was done using both the above ways and different sub techniques with different results. Simple NLP based method using sentence scores was used to extract top 5 sentences of the articles, it was observed that this method ignores the importance of nouns like names, places etc., and hence the output may not be meaningful.
TD-IDF based method was also used to derive the sentence scores and then extract top 5 sentences for the summary, results were impressive and all the nouns, places were given due importance however there was still a drawback that this approach ignores the order of words. Lex Rank and text Rank techniques were also used to summarize the text and results obtained had a good rouge score. The abstractive text summarization techniques use deep learning models to generate a summary like a human written summary. LSTM model was used to generate the same and the results obtained were not as expected which were attributes to the fact that RNN model requires high configuration machines to train the model which were not possible at this point of time considering the scope of the project.

***Keywords**: Text mining, abstractive text summarization, extractive text summarization, natural language processing, tf-idf, Python*

2.6 Telecom Customer Segmentation and Churn Prediction

In this project we present analytical models for two very prominent use cases prevalent in telecom industry i.e. Subscriber Segmentation & Churn prediction. The models thus created are based on proven statistical frameworks and include domain expert inputs at relevant stages of the model to give practical outputs and insights.
The analytical models are based on actual subscriber generated data points across various network and business support systems of a telecom operator. We have used 24 weeks of data collected over ~4000 subscribers across various services such as Data, Voice (On-net, Off-net), Roaming, SMS etc. and where required created relevant features from a combination of available variables. The data is mined using statistical techniques and specifically use widely accepted and used analytical techniques i.e.
− Clustering (K-means algorithm)
− Predictive analysis using classification algorithm (Logistic Regression algorithm)
The models are cross-validated against equally acceptable analytical algorithms for benchmarking the outputs. While creating the models in the first place, desired best practices of building right number of samples as well as transforming data to conform to the statistical requirements of the algorithms have been followed in letter and spirit to build robust models. Further, the models have been tested for their stability using unseen data (in case of churn prediction) and validation techniques (in case of clustering). As a result, we can present to business a clear view into four type of segments the cluster of subscribers throws up. We also have labelled the overwhelming characteristics displayed by the members of a cluster with respect to the type of services they consume whether Data, Voice, SMS, all or none (idle/dormant users).

***Keywords**: Telecom Industry, Subscriber Segmentation, Churn Prediction, Clustering, Predictive Analysis*

2.7 Predict Defaulters using Machine Learning

Credit Card and loan payment default is the one of the biggest issues faced by the retail finance industry. It has a cascading effect as increasing defaults are compelling these organisations to contract their credit card and loan portfolios immensely. The design of efficient fraud/default detection algorithms is key for reducing these losses, and more and more algorithms rely on advanced machine learning techniques to assist investigators. In this project, we have proposed the overall process of detecting the probable defaulters based on machine learning and process large amounts of financial data. F-measure and ROC curve are used to validate our proposed model. We also further cluster defaulters for further business information.

***Keywords**: Machine Learning, Payment Defaults, Financial Defaulter Detection, Supervised Method, Unsupervised Method, Feature Selection, SMOTE*

2.8 Stock Options Trading using Machine Learning

Stock markets are unpredictable as the fluctuations in the prices over time depend on several factors. This makes the investors full of doubts. However, if we look to the stock markets and behaviour of stocks prices in more detailed manner, we can observe patterns that define the generic evolution of the stock. Based on forecasted trends, we have to build our hypothesis for the analysis so that we will use Predictive Machine Learning Classifiers (e.g. Regression) with stock data that spans over a long-enough period of time, we can approximately predict the behaviour of the stocks and the evolution of their prices over time though Time Series Forecasting. After all our main objective is to invest in different stocks to get maximum profits for certain level of risks, for that we are creating a diversified portfolios of stocks with least standard deviations to minimize risks try to quantify that risk factors. To quantify the risk probability, we use Value at Risk (VaR) so that we will be able to analyse the downturn in the given

portfolio. Use of derivatives instruments like options (call and put) on a given stock help to hedge or maximize our return.

***Keywords**: Stock Markets, Forecasted Trends, Regression, Time Series Forecasting*

2.9 Upsell Lead Prediction of IndiaMART Clients

IndiaMART.com is a B2B Online Marketplace for Buyers and Sellers where Sellers showcase their products and Buyers search / browse the platform to find Sellers for their requirement. The clients (Sellers) are given Buy Leads through its Web enabled platform, enabling Buyers to find relevant Suppliers and vice-versa. This project attempts to solve IndiaMART's problem of identifying on-boarded Silver Clients on the platform who are the most likely candidates for upsell to Gold / platinum package, thereby maximising ROI for both client and IndiaMART. With the ready dataset, various algorithmic techniques such as Logistic regression, Random Forest, Combined Probability – Neural Net and Neural Net Ensemble were tried. The results were then presented to the business where business put forward further expectations to be able to predict the exact service which must be pitched to client. Business also demanded to know that gestation period for each client upsell.
Data with respect to the above requirement shall then be collected and further be presented to the business. For now, the model is being sent for implementation where at the start of each month, we will use this model to identify prospective clients which should be attempted upsell in the month.

***Keywords**: Online Marketplace, Logistic Regression, Combined Probability*

3.1 Exploratory Study to Understand Driver's Behaviour and Vehicle Condition

Driver's behaviour strongly impacts traffic security and performance of the vehicle. Aspects of driving such as over speeding, excessive engine idling, harsh braking has major impact on fuel consumption and road safety. Telematics devices plugged in vehicle can play a beneficial role in promoting safe driving thus reducing the accidents and their associated costs. Moreover, insurance companies are now trying to adopt UBI (usage-based insurance program), which calculates the premium based on the driver's driving behaviour and time of driving because reckless driving badly affect the profitability of insurance companies. Therefore, it is essential to have a system that can accurately profile a driver based on his/her driving behaviour using both real time and long-term data, as well as other related vehicular and environmental data. Telematics can also bring about benefits for tracking Vehicle health by the close monitoring of different aspects of vehicle performance, including engine malfunctions, faults or warnings, thus avoiding vehicle downtime. Thus, the problem this project tries to address can be defined as follows:

- How to profile a driver based on his/her driving behaviour using real time and long-term data?
- How to analyze whether vehicle needs maintenance or not?

***Keywords**: Traffic Security, Over Speeding, Harsh Braking, Telematics, Insurance Program*

3.2 Cancer Diagnosis Transformation

Purpose of this report is solely for academic. Selection of healthcare as domain was based on considering the significant contributions that can be provided by machine learning. The accelerating power of machine learning in diagnosing disease and in sorting and classifying health data will empower physicians and speed-up decision making in the clinic. Cancer is the second leading cause of death globally; early detection is crucial and can go a long way in saving lives. We have tried to automate the manual classification process done by the pathologist, wherein cancer variations are classified into mutation classes. The key role of pathologists is to examine the report to determine precise type and

severity of the disease. Pathologists are integrating AI and machine learning software into their workflows to cut down work time, reduce misdiagnoses and connect with fellow pathologists–in general, making their departments more cost-effective and high functioning. Over the course of a cancer patient's diagnosis and treatment, pathologists record highly descriptive and specific observations of cells and tissues in pathology reports. It is meant to serve as a tool to save pathologists' time by instantly highlighting potentially problematic areas of the scan, instead of having to do this manually. We have focused to work towards enhancing the work of pathologists by building a model to classify the genetic variation based on research literature into one of the 9 mutation classes. We have obtained dataset from ''Memorial Sloan Kettering Cancer Center (MSKCC)" website, which is one of the world's oldest and largest private cancer center that provides exceptional patient care and does lot of innovative research on cancer. We have built 5 different classification models and 1 ensemble model and obtained outputs for genetic variation classification into one of the 9 mutation classes with multiclass log-loss as our key performance metric and created graphs in Python and Tableau for visual interpretation.

***Keywords**: Healthcare, Physicians, Cancer, Pathologists, Mutation, Python, Tableau*

3.3 Sales Promotion Model for Big Bazaar

Retail chains world over spend a substantial part of their marketing budgets on promotions. While promotions account for a significant share of retailer revenues, promotion intensity continues to increase. Despite substantial investments in promotions, retailers often have little understanding of the true performance and thus struggle to determine which promotions are working and which are not. Future Retail runs various loyalty programs and festive offers which provide their customer more opportunities to avail discounts. Customers can use these offers or loyalty program to either avail discount or to make payments.
This project is aimed to analyse customer buying patterns using the data collected during the time spanning 01/01/2016 to 06/30/2017. Eventually evaluation of performance of historical promotions run and present recommendations on appropriate promotions to increase wallet share of customers would be the successive step.
There are two ways to measure effectiveness:
1. Scale of efficiency—the extent to which cost of promotional spend is minimized for a given number of promotions availed by customers
2. Scale of monetary improvement—the extent to which spend on promotion reaps/achieves an increase in sales or profit
The project is aimed to evaluate effectiveness of promotions on monetary improvement and thereafter to recommend improved promotional schemes to further improve the sales.

***Keywords**: Retail Chains, Retailer Revenue, Promotion Intensity, Loyalty Program*

3.4 Credit Risk Analytics of Peer to Peer Lending Models

Credit risk is a serious problem in banking and is of concern to a variety of stakeholders -institutions, consumers and regulators. It has been the subject of considerable research interest in banking and finance communities and has recently drawn the attention of statistical researchers.
Credit risk is the risk that the borrower from a bank will default on the loan and/or the interest payable, i.e. that it will not perform in terms of the conditions under which the loan was granted. This is damaging to the bank, not only because of the actual loss eventually incurred, but also in terms of the time that management and bank counsel expend on attempting to recover the loss or a portion of the loss.
The agenda of this project is to create multiple Credit Risk models using multiple techniques such as regression and random forest try to develop and identify the best model that could predict the probability of default with maximum accuracy and with least false positive rate. The model developed through study would benefit the Banking companies to predict PD with higher degree of confidence level

and consequently would lower the value of NPA of the bank. With a healthy banking sector, and the loans reaching to the right consumers the banking sector, the economy of the nation as well as the consumer would prosper.

***Keywords**: Credit Risk, Regression, Random Forests, Probability, Banking Sector*

3.5 Revenue Maximization in Retail Stores through Effective Campaign and Personalised Recommender Systems

In this project, we explore the problem of purchase cart recommendation in the field of brick & mortar retail stores. Recommender Systems application is widely used in online sales platforms and content providers for e-commerce, movie recommendation etc. We want to explore usage of recommender systems in brick & mortar stores while keeping in mind their operational constraints, specifically restricted interactions between customers and products. Physical shops have to cope with new competitors, e-commerce actors like Amazon, Flipkart & others, who have redefined ways customers purchase - faster, cheaper and often more targeted. Large physical retail store chains have started to compete on same ground, exploring online selling portals of their own. But, major chunk of their sales is still from physical stores. We are proposing to use recommender systems to provide customized services to their customers, thereby improving customer satisfaction, enhance their loyalty and thus increase their benefits. One possible application of such a strategy would be to use personalized recommenders to provide purchase cart recommendations, rather than individual product suggestions. We studied different types of recommender systems and decided to implement a customized recommendation engine strategy to achieve our goal. We expect our solution to bridge the gap between physical stores and online platforms as far as customized customer experience is concerned and act as a valuable brick & mortar retail store campaign strategy.

***Keywords**: Recommender Systems, E-Commerce, Customized Services*

3.6 Anti-Fraud Bot for Bajaj Allianz Life Insurance

The main objective of this project is to develop a chatbot that can help the customers in answering the queries instantaneously around the clock without having to wait for a human to assist them on their queries. In the financial service industry, customer satisfaction is as equivalent or even more important than to acquire a new customer. And also it is difficult for the internal employees to remember the policies and procedures followed by the organisation. Our chatbot will be the answer to both the crisis. With the help of deep learning, the chatbot is designed intelligently to cognizant itself from its past experience and also redirect to the users to the correct point of contact if it fails to recognize the question. That being said, more the questions asked, wiser will become the chatbot. Demand for chat bots that imitate human conversations for solving various tasks is steadily increasing. Chatbots are the new holy-grail of augmenting customer engagement and brand presence. With the help of intuitive chatbots, insurance companies are able to explain complex products to their customers, drive brand engagement, and improve sales and distribution with minimal human effort and intervention.

***Keywords**: Chatbot, Deep Learning, Customer Engagement, Brand Presence*

4.1 Analysis of Claims Data for Invest Assure

Purpose – To provide a brief analysis of Invest Assure endowment policy claims management using claims data of policy holders provided by Bajaj Allianz Life Insurance Company Limited (India). Design/methodology/approach – For convenience, this analysis of claims data is somewhat arbitrarily classified by topics as follows: age-based claims analysis, demographic analysis, smoking habit based

claims analysis, estimation of distribution of claims using six-year data, develop a model to screen the future prospects (Customers) to approve the policy or not and prediction of future claims.
Findings – Emphasis is placed on a discussion of different kinds of characteristics affecting the claims on a comparison of the statistical models and methods used to analyse such data. Prediction of future customers who might claim is also performed using the model.
Project limitations/implications – Since the literature on claims management data is vast, more work on this problem is needed. Due to data limitations, in-depth analysis on reasons of claim, connecting it with mortality rate, and other such deep-dives weren't feasible.
Practical implications – This report points out how claims management can be made more efficient. This analyses can be performed and applied for all insurance types and across the insurance industry not keeping it specifically limited to Bajaj.
Originality/value – The report reviews different statistical models and methods used to analyse claims data. The statistical models and methods presented are valuable and meaningful for claims management analysis and future claims prediction.

***Keywords**: Endowment Policy, Demographic Analysis, Claims Analysis, Statistical Models*

4.2 Financial Analytics on Non-Performing Assets (NPA) in Banks of India

The topic of the study is "Financial Analytics on Non-Performing Assets [NPA] in Banks of India". This analytics engagement is carried out for the Banking and Financial Services Industry in India in general, and not for any specific Company or Financial Institution. This study is conducted on the NPA of commercial banks for the period 2007 -2017. This study helps the industry, non-banking financial institutions to benchmark and the investors and businesses to understand the behaviour of NPA for making right business decisions. The study analyses the NPA data and identify whether the NPA ratio is same or different across commercial banks using ANOVA technique. The study also creates clusters using K-means clustering algorithms and predicts the cluster group using Linear Discriminant Analysis. The final step of the study forecasts NPA across commercial banks using multiple linear regression model.

***Keywords**: Financial Analytics, Non-Performing Assets, ANOVA Technique, K-means Clustering Algorithms, Linear Discriminant Analysis*

4.3 Predicting Trade Price of the Bond

Bond prices are a result of complex markets, decisions and polices, and predicting the future price are difficult due to dependencies and complexities. Wide gap between the amount of relevant information available to trading equities verses corresponding trading corporate bonds, makes the task challenging. Predicting the bond price becomes even more challenging due to the lack of full disclosed information, and accuracy is not the only consideration–in trading situations, time is of the essence. Hence machine learning in the context of predicting bond price should be fast and accurate.
The goal of this project is to develop a global model using the techniques and algorithms of machine learning and a set of data describing trade histories, intermediate calculations, and historical prices made available by Benchmark Solutions, a bond trading firm, in order to accurately predict bond prices. Our aim is also to paying more attention to trading dynamic, various features and structure of the bond.

***Keywords**: Bond Prices, Future Price, Machine Learning, In-Trading*

4.4 Predicting Long Term US Treasury Yields

This study discusses the various factors which potentially influence the evolution of the long-term US treasury yields. This has been an area of interest across the financial sector given the advent of number

of hedge funds, investment banks, mutual funds etc. who are very active in the US interest rates market from a trading perspective. Additionally, long terms US yields are one of the key determinants of mortgage rates in the US. US Mortgages and securitized products market is one of the biggest markets in the world. Given the influence of these yields in the pricing of mortgages, the evolution of these yields is not only interesting for speculators but also for hedgers who have exposure to these markets like mortgage originators and servicers.
The study builds and compares various models using the variables which are assumed to potentially impact the evolution of the US long terms yields and recommends the best model that can be used to predict the direction and the extent of the future move in these yields. Lastly, the model also highlights some of the challenges that the users should be aware of when using these models to make trading decisions.

***Keywords**: Treasury Yields, Financial Sector, Hedge Funds, Investment Banks, Mutual Funds, Mortgages*

5.1 RAG Classification for Measuring Performance of Life Insurance Agents

The agent/broker channel in the insurance sector plays a prominent role in the value chain. They are very influential in determining the premium revenue, brand value, customer retention and customer service experience. Insurers that can manage their agencies effectively are better positioned to achieve a competitive advantage and hence the analysis of agency performance becomes crucial for an insurance provider. The performance of the agents can be done on various parameters such as profitability, product portfolio, customer retention as well as different servicing aspects of customer relationship management such as complaints management, documentation, servicing customers based on requirement. Analytics can help identify the good performing agents and provide them the necessary support to perform better. It can also help in incentivising the good performers and identify good practices that can be shared across other agents. Analytics can also help in understanding those agents that are consistently low performers and either take corrective actions based on root cause analysis or prune such agents. At Bajaj Alliance Life Insurance Company (BALIC), seven variables have been identified for performance measurement on the customer servicing front. The performance of the agents is compared with pre-defined threshold values for each variable and a final score is arrived at to classify the performance as Red, Amber or Green. BALIC identifies the agents classified in the Red category and communicates areas of improvements that they can work on to improve their performance.

***Keywords**: Insurance Sector, Premium Value, Brand Value, Customer Retention*

5.2 Behaviour Scorecard

The organization provides credit to corporates. The objective of the project is to build a behaviour scorecard to identify which of its live customers are likely to roll forward to higher delinquency bucket in next few months, from which it is not possible to recover.
Two approaches were taken for data preparation. In the first approach, every 3rd transaction of the customer was sampled for modelling. In the 2nd approach, a 3-month rolling performance window was used to sample 1 transaction from each customer- thus maintaining the uniqueness of customer. For each sampled transaction, historical data of previous 12 months was taken for observation of behaviour ("observation window") and next 4-6 months was taken to measure the performance ("performance window") of the customer. The resultant dataset was highly imbalanced, with just 3% (1.5% in 1st approach) of the transactions representing the target class to be predicted. SMOTE technique, used to bump up the minority class transactions, was optimized a 15% to give the best results. Logistic Regression algorithm was selected as Business wanted to know the impact of the variables on the outcome. Due to high multi-collinearity & large no. of features (72), model was run iteratively.

***Keywords**: Behaviour Scorecard, Delinquency Bucket, SMOTE, Logistic Regression*

5.3 Stock Market Analyst Rating

Analysts have been actively evaluating companies as long as there have been stocks, but they're more popular and get more exposure than ever thanks to round-the-clock stock market news and online resources. Some analysts' notoriety has also increased. But while analysts typically have similar credentials, they aren't all the same.
This project will help to differentiate among different stock analyst as per the recommendation given by them. We have used the Machine Learning algorithm to create a framework which provides rating to market analyst on the scale of 1 to 5 on the basis of their past performance.
Although a disclaimer that past performance is not a predictor of future performance accompanies this promotional material, there is a strong implication that past performance is a good measure of the market analyst's skills.

***Keywords**: Stock Market, Shares, Financial Markets, Rating System, Predictive Analytics, Data exploration, Data Cleansing, Classification, R, Excel, Random Forest, Naïve Bayes, Decision Tree, Machine Learning, Market Capitalisation*

6.1 Factors Influencing Infant Mortality Rate

This project is to identify the key factors that affect the Number of Infant deaths reported in India. The Variable Importance is determined, and Infant Mortality is predicted by using below Machine Learning techniques:

- Random Forest
- Linear Regression (using Principal Components)
- Neural Network

Apart from this, using the unsupervised learning algorithm (k means), to classify the data based on the heterogeneity in the data points.

***Keywords**: Variable Importance, Infant Mortality, Machine Learning*

6.2 Customer Segmentation and Profitability Analysis (for a Cash Management Company)

Cash Management and Logistics is an important segment of market in countries like South Africa. Although Africa can provide great opportunities, a treasurer looking to run operations needs to be cognizant of the risks and challenges associated with the region. We have partnered with one such cash management company SBV Services Pty. Ltd. headquartered in Johannesburg, South Africa, to analyse their customer segments and produce to them the key factors that are affecting their profitability within each customer segment.
One of the important business problems of SBV was that they had various types of customers who have subscribed to different types of services from SBV, however, the firm was not able to understand which customer segment is profitable and the factors affecting the other customer segment that less profitable. We collected the data pertaining to the customer details, customer type details, and transaction type, vehicle run time, revenue, direct and indirect costs which will determine the profitability of the transaction. The data was then subjected to a K-means clustering technique after obtaining the optimal number of clusters through an elbow curve. The outcome of the analysis was that the company had 4 distinct customer segments of which two clusters were profitable and the other two were making loses. The important factors that were affecting the profitability are service time of each cluster, indirect cost, direct cost spent towards the staff which were dependent on the type of the customer viz. wholesale or retail.

***Keywords**: Cash Management, Logistics, Customer Segments, K-means Clustering Technique*

7.1 Sentiment Analytics for Stock Market, Crypto, FX and Commodity

The purpose of this work is to observe how well the changes in prices of financial product, the rise and falls, are correlated with the public opinions being expressed in tweets about the respective underlying product. Understanding author's opinion from a piece of text is the objective of sentiment analysis.
In this project, we would like to use applied sentiment analysis and supervised machine learning principles to the tweets extracted from Twitter and analyse the relation between respective market movements (viz. Stock Indexes, FX, Commodity & Crypto) and sentiments in tweets.
At the end we will try to find out whether there is a strong correlation exists between the rises and falls in financial product prices with the public sentiments in tweets.

***Keywords**: Sentimental analytics, Text mining, Sentiments, Machine Learning, Twitter, Sampling, Trend Analytics, Tidy Text*

7.2 Predicting Cervical Cancer Likelihood and Indicators using Machine Learning

Cervical cancer is a type of cancer that occurs in the cells of the cervix. Various strains of the human papillomavirus (HPV), a sexually transmitted infection, play a role in causing most cervical cancer. One can reduce your risk of developing cervical cancer by having screening tests and receiving a vaccine that protects against HPV infection.
To identify and confirm the presence of cancerous tissues in the body, various screening tests are conducted. Screening means checking one's body for cancer before one has symptoms. As in many other diseases, the existence of several screening and diagnosis methods creates a complex ecosystem from a Computer Aided Diagnosis (CAD) system point of view. Hence, there is need to predict individual patient's risk and the best screening plan during her diagnosis.
The objective of this project is to use Computer aided diagnosis and machine learning techniques to classify respondents into Cervical Cancer or healthy category based on multiple test results provided in the dataset thereby preventing the need of doing multiple screening tests. As the data is related to cancer, our project team would need to have working knowledge of the relationship between cancer genes and cancer progression. As 2 of our team members are in the pharma industry, we have access to microbiologists and researchers to help us here.
The primary objective of this project is to avoid the above uncomfortable and painful situations by predicting risk of cancer using historical data and do early detection without the need of doing multiple screening tests. This reduces the burden of healthcare cost and patient compliance. The application of the model is for both patients and clinics. For patients, it would be early prediction of cancer without undergoing any painful screening tests. Only high-risk patients must undergo for screening tests for confirmation. Early diagnosis (Stage 1 or Stage 2) will have more therapeutic options for patients with much better outcomes. Our approach is to consume the data collated from surveys on the lifestyle habits, medical history etc. from available sources conduct analysis using machine learning techniques and conclude with a model that

- Helps to predict the likelihood of cancer without doing the different types of painful tests
- Identifying the parameters that increases the likelihood of cervical cancer

***Keywords**: Cervical Cancer, HPV, Computer Aided Diagnosis, Machine Learning, Pharma, Healthcare Cost, Patient Compliance*

7.3 Football Player's Marketability Index

Football is the most popular sport in the world. It is played by almost 211 countries. Its popularity has increased even more by the rise of club football. Clubs like Real Madrid, Barcelona, Manchester United, AC Milan, Juventus, Arsenal, Bayern Munich and many more have huge fan bases. Teams continue to form part of the cultural and sentimental heritage of cities and national teams, which continue to arouse

passions. They can be considered another national symbol. The fan base is so huge that it has gone way beyond just the locals supporting the club to people from all across the world supporting it. With the advent of Broadcast TV and Social Media, the game has become even more global. Football has changed and so has its economics and marketing. Now, clubs buy big money players for two reasons. One is obviously to make the club better by having someone with specific skill set and the other reason is to help the club in marketing which in turn helps them in increasing their revenue, which helps them in expanding their base to different countries. Paris Saint-Germain (French Club) bought Neymar from Barcelona on a record shattering transfer by triggering the release clause of $263 million. On the back of his transfer, there was plenty of focus on shirt sales. Around 120,000 Neymar PSG jerseys were sold in the 30 days that followed his switch—raising $10.1 million. Commercial value may be hard to judge, but one thing that is more easy to monitor is social media numbers, which can measure how influence and business is growing. Across Instagram, Twitter and Facebook, PSG's followers at the end of the 2016/17 season totalled 38.1 million. That figure soared by 5.3 million ahead of Neymar signing his new contract. Now, 12 months on from the transfer, it is apparent from the figures on PSG's accounts that followers have soared by a further 11.8 million across the three social media platforms. Take into consideration also that Neymar has a personal following of 165 million and his reach to fans, businesses and consumers across the world becomes hugely significant. Even in the Indian Super League, when the league started, every team signed a marquee player. Players such as Luis Garcia, Alessandro Del Piero, Freddie Ljungberg and David James were signed by different clubs to attract the Indian audience towards the league by marketing all these players before the start of the season. After the season ended, clubs were asked to retain 2-3 players for the next season. Kerala Blasters retained CK Vineeth and Sandesh Jhingan. Why these two? Sandesh Jhingan was retained for being a good defender in the previous season but Vineeth was retained for being the local boy which helped Kerala in their marketing.

***Keywords**: Machine Learning/Deep Learning, Sentiment Analysis, Player Rating System*

8.1 Loan Approval Prediction based on Parametric, Non-Parametric and Tensor Flow Machine Learning Models

Due to easy availability of loans, in today's world, taking loans from financial institutions has become a very common phenomenon. Everyday a large number of people make application for loans, for a variety of purposes. But all these applicants are not reliable, and everyone cannot be approved. In the past as well as this year, we read about a number of cases where people do not repay bulk of the loan amount to the banks due to which they suffer huge losses. The risk associated with making a decision on loan approval is immense. Today's era is all about "Data Management". Nowadays, banks have included a large amount of information in its evaluation of loan issuance, and some of this information has a vague causal relationship with the loan default rate. The growing amount of data due to improved data capture and data storage technology has bring us to a new perspective on this problem which is used to be accomplished by financial and economic analysis. Accurate prediction of whether an individual will default on his or her loan, and how much loss it will incur has a practical importance for banks' risk management is very important. So, the idea of this project is to:

• Study to whom the loan was granted before and on the basis of these records/experiences and extract patterns, which would help in predicting the likely defaulters, using classification data mining algorithms.

• Banking Big Data are stored in form of clusters, so understand the distributed environment and suggest a suitable ML algorithm to work in distributed scenario.

***Keywords**: Loans, Financial Institutions, Risk, Data Management, Economic Analysis, Data Mining Algorithms, Machine Learning*

8.2 Develop the Credit Scorecard for Lending Club

As a part of project, we studied the problem faced by the P2P lending industry, where lender can earn higher rate of interest, whereas borrower can spend lower amount of interest as compared to the conventional banking system, where the concern is how risky is the borrower & given the borrower risks, should we lend him/her or not. After the thorough study we conducted, we were able to suggest a credit scorecard for the Lending Club Corp., & explored various machine learning classification model techniques for predicting the default rate of customers like Logistic Regression, Random Forest, Lasso & Ridge & Ensemble & cross validated our data on the test data set.
Following deliverables will be met to address the problem faced by the Lending Club-

1. Analysis the loan payment dataset of the Lending Club Corp, to better understand the best borrower's profile for investors.
2. Perform segmentation of the loan database into finish cases & current outstanding loans.
3. Breakdown the composition of the default cases.
4. Examine the correlation between default cases & the independent variables/indicators.

Finally, the probability of the customer defaulting on the loan pay out was calculated for each individual customer and then based on that the loans were classified into the likelihood of high, medium and low risk of delinquency. The low risk customers should be given rebates in the interest rates which can improve customer retention rate. The Machine Learning Algorithm or an analytical model can be further extended to analyse the data sets available from all P2P lending companies available in the Market.

***Keywords**: Machine Learning, Logistic Regression, Random Forest, Segmentation, Correlation, Probability, Likelihood*

8.3 Accent Analytics

Accent is a distinctive way of pronouncing a language, especially one associated with a particular country, area, or social class. English is very close to becoming a global language with an estimated 1.5 billion speakers globally. Yet, it remains a fascinating subject because it is spoken in different ways across countries. Among the many issues faced by Automatic Speech Recognition (ASR) systems, effectively handling accents are one of the most challenging. Particularly when working with languages that have highly varied pronunciations such as English, an ASR system trained on only one accent might only be effective for a minority of the speakers of that language. The Objective of this project is identifying and accurately labelling English accent from different geographical regions. There are many accents given the widespread use of the English language. Therefore, we limited ourselves to the four accents: American English, Arabic, Mandarin and Indian. The main objective is to design and train a model to identify and label these four accents in spoken English.
Primary Objective of the study are:
• The approach to this task of accent classification consists of feature extraction and developing machine learning classifiers
• Extract acoustic feature MFCC (Mel Frequency Cepstral Coefficient) as an input to machine learning algorithms
• Build various Machine Learning multi classification algorithms to find the best fit model
• Predicting / Classifying short 30-second accented voice clips as one of target Accents
The Audio recordings of people with different Accents are scrapped from the Speech Accent Archive an online repository of spoken English. Also, Speakers representing different native languages read a common paragraph in English was recorded.

***Keywords**: Accent, Geographical Regions, Classification Algorithms, Machine Learning, Predicting*

8.4 Product Analytics for Retail Brand Apparel_Alhassina

The projects would contain analysing the Brands & Apparel_Alhassina dataset and implementing product analytics. The models have been built after doing missing value treatment, outlier treatment, EDA, Univariate and Bivariate Analysis on the data. The 3 main objectives that we have work towards in this project are as listed below-
Sales Revenue Prediction: A sales forecast is an essential tool for managing a business of any size. It is a month-by-month forecast of the level of sales you expect to achieve.
Model Used – Linear Regression Model and Accuracy Check Commonality of product: Based on customer characteristic which product is commonly purchased need to pitch the product to the customer based on customers purchasing habits.
Model Used – Market Basket Analysis and Association Rules Customer segmentation: Identify customers who are high/low spending customers and who are engaged with you and find which market has highest consumers.

***Keywords**: Product Analytics, Missing Value, Outlier, Univariate and Bivariate Analysis, Sales Revenue Prediction, Linear Regression, Market Basket Analysis*

9.1 Travel Time Prediction and Cab Requirement Forecast for a leading Transport as a Service Organisation

The main objectives of this project are:

- Predict and benchmark the travel time from different zones to MNC office and vice versa in order to manage expectations of the employees using a regression model
- Forecast a range of weekly Cab requirement for login and log out shifts per zone and per vehicle type (4 seater/6 seater) on the basis of historic data using a time series model

Keywords: Regression Model, Forecast, Historic Data, Time Series Model

9.2 Hybrid Model for Solar Power Forecasting

Integration of solar energy into electric network is one of the most promising approach to meet the ever-increasing demand for energy, without significantly impacting the environment. To ensure efficient operation of power generation and distribution systems, it is essential to have reliable forecast information of solar resources. Forecasting enables quantification and estimation of the available energy, and in-turn resulting into optimal management of transition between intermittent and conventional energies. In this study, an attempt has been made to explore and evaluate different approaches for forecasting the solar power generation for the next day. The approach developed in this study is based on the simulations from Numerical Weather Prediction models (NWP). NWP uses mathematical models of the atmosphere and oceans, to predict the weather using information about current and past weather conditions. Results of these simulations are used as inputs for statistical model, estimation of bias and its corrections, and finally to create models for forecasting the power generated from a solar power plant. This very short-term forecasting service is very important for grid operators, it ensures the stability of the grid, and aid the power plant operators to keep the plant deterministically controllable. Several different forecasting techniques were investigated, and the most appropriate method for forecasting the daily solar irradiance and the amount of energy produced by the plant was selected. Bhadla Solar Park, Rajasthan, India, has been considered for evaluation of the prediction model presented in this case study.

***Keywords**: Forecasting Solar Power Output, Simulating Weather Parameters, Bias Correction, Grid Optimization, Solar Power Integration*

9.3 Youtube Video Analysis

Analysis of structured data has seen tremendous success in the past. However, analysis of large- scale unstructured data in the form of video format remains a challenging area. YouTube, a Google company, has over a billion users and generates billions of views. Since YouTube data is getting created in a very huge amount and with an equally great speed, there is a huge demand to store, process and carefully study this large amount of data to make it usable. Today, YouTube is the largest user-driven video content provider in the world; it has become a major platform for disseminating multimedia information. A major contribution to its success comes from the user-to-user social experience that differentiates it from traditional content broadcasters. This work examines the social network aspect of YouTube by measuring the full- scale YouTube views graph, comment graph, and video type corpus. There are a lot of companies that are actively creating content on YouTube aiming to reach their user bases. The content is posted in their respective channels. Content is either served directly on the channel or by pushing it on other popular videos as advertisements. Companies have varying strategies to make their content popular; some go the organic route while other are big on advertisements. The report documents the effort to a data driven approach for understanding the YouTube market by taking client data of 35 channels. Through the historical video content, efforts will be made by applying different algorithms: Neural Network, Logistic Regression and Random Forest to understand the accuracy of the algorithms in predicting how much of a channel content gets hits through Advertisements. Time-series analysis is performed to study the relationship between key variables and their variability. Finally, Multilevel Regression analysis will be performed to understand the effect of these parameters and accuracy these parameters would lead to. All the analysis and the findings will be delivered through statistical software R.

***Keywords**: Multimedia, Comment Graph, Video Type Corpus, Neural Network, Logistic Regression, Random Forest, Time Series, Regression Analysis*

9.4 Bank Marketing

Recently, economic depression, which scoured all over the world, affects business organizations and banking sectors. Such economic pose causes a severe attrition for banks and customer retention becomes impossible. Accordingly, marketing managers are in need to increase marketing campaigns, whereas organizations evade both expenses and business expansion. In order to solve such problems, data mining and analytics techniques are used as a factor in data analysis, data summarizations, hidden pattern discovery, and data interpretation. In this project, various data analytics techniques have been implemented, using a real marketing data obtained from Portuguese marketing campaign related to bank's term deposit subscription. The project aims to improve the efficiency of the marketing campaigns and helping the decision makers by reducing the number of features, that describes the dataset and spotting on the most significant ones and predict the deposit customer retention criteria based on potential predictive rules.

***Keywords**: Economic Depression, Banking Sectors, Marketing Campaigns, Data Mining, Hidden Pattern Discovery, Customer Retention*

9.5 Out brain Click Prediction

The internet is a stimulating treasure trove of possibilities. Every day we stumble on news stories relevant to our communities or experience the serendipity of finding an article covering our next travel destination. Outbrain, the web's leading content discovery platform, delivers these moments while we surf our favourite sites. Currently, Outbrain pairs relevant content with curious readers in about 250 billion personalized recommendations every month across many thousands of sites.

In this Project, our objective is to predict which pieces of content their global bases of users are likely to click on.

***Keywords**: Travel Destination, Content Discovery Platform, Prediction, Global Basis*

10.1 Consumer Loan Analytics

The key objective of the project was identifying variables and consumer segments that will allow lending to larger audiences and simplify the current lending process.
This was a challenging project at multiple levels given the big data dataset spread across multiple variables and no access to client for clarification/ reasoning or professional infrastructure. However, overcoming the challenges, team has extracted a data subset 2 which has been used for this project. Data modelling has been first done by using Logistic Regression and later by Random Forest technique. These techniques were identified as best suited techniques given the nature of dataset (independent, dependent variables and binary output) and the business objective being exploratory in nature. Multiple tools like R, Python, and Cloud were used for data merging and modelling. With Logistic Regression team was able to identify 55 significant variables mainly spread across Demographics and prior Lending history like amount, credit period, future instalments, etc. This model had accuracy of 90.5%. These findings can be further refined using neural network and XGboost model. It is possible that there are certain findings that might come across as generic in nature; the final recommendations need to be discussed and agreed with the business team for implementation. Given the generic nature of findings, they can be implemented by financial companies across board and not just 'Home Credit'. Business teams need to understand the geographical nuances and legislative restrictions before implementation of the findings.

***Keywords**: Customer Segments, Data Modelling, Logistic Regression, Random Forest, Neural Network*

10.2 Health Economics and Outcome Research of a Public Health Insurance Organization

While India is passing through demographic and environmental transition, the landscape of healthcare is witnessing constant addition of disease burden, putting strain on individual health as well as public health resources. Total healthcare expenditure in India is estimated to be Rs. 5.3T, amounting to 3.8% of country's GDP. Government alone spends nearly Rs. 1.6T on several healthcare initiatives. Targeted healthcare can help reduce this burden, but administrators at healthcare organizations need to develop data-driven strategies to achieve this. In the current study 5-year data of a particular public health insurance organization is analysed to provide insights on patient attendance in the clinics, overall expenses of the Organization, patient profiling on the basis of prescriptions & treatments, hospitalization claim pattern and resulting financial burden. Patient attendance patterns were studied in the clinics between 2014 and 2018, and short term forecast for next 6 months ahead was generated using time series forecasting. It is observed that the increasing attendance in the clinics has led to 70% increase in the median waiting-time of the patients over the past 5 years. G/G/m Queueing model was constructed to estimate patients' waiting-time in three clinics. Suggestions are made to improve the waiting-time of the patients by reducing non-clinical activities and optimizing number of doctors in each clinic. Reduction in waiting-time can significantly improve patients' satisfaction, while increased availability of doctors would improve the quality of diagnosis and prescriptions. Expense data analysis showed increasing trend in expenses under most categories. Category-wise expense forecast was generated for next two financial years, which estimated total expenses to increase from Rs. 1.07B in FY 2018-19 to Rs. 1.23B in FY 2019-20 and Rs. 1.35B in FY 2020- 21. This would help the Organization plan and propose budget for coming cycles. Patient profiling based on prescribed medicines and hospital treatments showed significant increase in certain clinical conditions. Anti-diabetic drug prescriptions have doubled, and emergency hospitalizations due to Respiratory ailments have increased 90% over the past 5 years. This information can be used for targeted/preventive care and planning special programs by the Organization to improve

beneficiaries' health. Financial implications of the Organization's medical expenses on the public funding was studied, and Insurance Claim Ratio (ICR) was calculated. The study showed ICR of 637%, which is 6 to 10 times higher than any private or public sector health schemes operating in India. Recommendations are made on limiting the scheme's ICR, which in turn will reduce burden on public funding.

***Keywords**: Healthcare, Prescriptions, Treatments, Claim Pattern, Time Series Forecasting, Preventive Care*

10.3 Churn Management

All industries suffer from churn – the loss of customers to the competitor. The survival of any business is based on its ability to retain customers. This is particularly true for Telecom Companies where cost of customer acquisition and retention is very high. The advanced analytical techniques can help CSPs reduce the Churn by proactively identifying potential churners and taking timely retention measures.

***Keywords**: Customers, Competitor, Retention, Telecom Companies, Customer Acquisition, Analytical Techniques, Churn*

11.1 Sentiment Analysis for Customer Service in Banks

The purpose of this report is to is to improve customer experience by utilizing AI and machine learning to develop an algorithm where customer E-mails are scanned, analyse the sentiment from the body of the message and automate customer e-mail categorization and prioritization for the banking sector. The main goals are stated below:
(1) Collect bank query related E-mail data, ranging from general information, escalations and requests
(2) Develop a machine learning algorithm that can perform text mining and sentimental analysis
(3) Provide priority based categorized information for management to prioritize and improve customer Service

***Keywords**: Customer Experience, Machine Learning Algorithms, Automation, Banking Sector, Text Mining*

11.2 Patent Based Technology Landscape Analysis and Grant Prediction for Trading Strategy

Investing in the right company has always been a challenging problem for investors. While large financial data sets and indicators are available, there may be implicit value drivers that are not being considered. One of the key value drivers especially for product based companies is the patent portfolio. In our view, a deeper analysis of the patent portfolio and its value will provide investors with an excellent tool to understand the growth of the company and drive investment decision making. Using a few large electronics companies, we compare how the business performance in terms of revenue and stock price is impacted by the patent portfolio of the companies as well as R&D expenses. We find that companies that have deep investments in patents like Samsung and Apple have a statistically significant correlation between revenue and stock price to the number of patents filed and granted. Given that patents are an important indicator of business performance; we proceed to formulate a trading strategy that is based on a forecast of the number of patents applied by the companies. Given the large number of patents being filed and the high cost associated with the same, companies need an analytics based tool to decide on patentability. Using machine learning models and NLP techniques, we predict whether a given patent will be granted or not. This will help the companies use the model as a dipstick before applying for patents. Given the wealth of information available in patents, an important opportunity for company leadership is to understand how the technology landscape is changing in the industry as well as for competitors. We use NLP based analysis on patent abstracts to determine how the technology landscape is changing over the years and where fresh investments are happening now. The purpose of all of the above patent analytics is to provide enabling tools for both the investor and an interested company's leadership.

***Keywords**: Patent Analytics, Trading Strategy, Grant Prediction, Hypothesis testing, Technology landscape, Machine Learning, Natural Language Processing, Latent Dirichlet Allocation, Legal Analytics, Time Series Forecasting, Word Cloud*

11.3 Analysis of Retail Store Sales using Machine Learning

A major challenge for large retailers is to address the needs of the consumers more effectively on a local level, while maintaining the efficiencies of central distribution. As the demand for mass customization by consumers grows, methods focused on store level optimization increase in value. The recent development in information technology and ubiquitous computing makes it feasible to move recommender systems from the ecommerce realm, where they are widely used, to retail stores. The world is moving towards online shopping due to the benefits it provides to the customers. Big e-commerce sites like Amazon, Flipkart etc. provide tailored recommendations to their customers based on their previous behaviour. These aid the customers during their shopping and provide satisfactory shopping experience. These recommendations are given by recommender systems employed by the sites. As these systems evolve and become more complex, more accurate predictions and recommendations are given to the users. If this continues, retail business might wane in the future. Hence, a system which uses cash receipt data to provide personalized recommendations to customers can help provide business advantage to the retail stores.

***Keywords**: Retailers, Central Distribution, Information Technology, Accurate Predictions*

11.4 Financial Risk Modelling for Corporate using Machine Learning

The credit risk defaults by companies have been increasing at an alarming rate which may cause huge unexpected monetary loss to the banks or financial institutions. This may directly affect the functionality of the financial institution to a great extent. In the above context, the analysis of the risk profiles of the companies becomes more significant. Banks or any financial institution should be able to predict the companies which are likely to default through the risk profile analysis and avoid lending to those companies. The primary objective of this project is to analyse the past and current financial data and predict the future risks. The process of analysis deals with categorizing the companies into one of the three risk profiles namely, low, medium and high. The project uses the following 6 machine learning algorithms to predict the risk profile of the companies.

1. Support Vector Machine
2. Linear Discriminant analysis
3. CART
4. Random forest
5. K-Nearest Neighbour
6. Naïve Bayes

The project uses the following parameters to assess the performance of the models.

1. Accuracy
2. Average Sensitivity
3. Average Specificity
4. AUC

***Keywords**: Credit Risk, Financial Data, Machine Learning Algorithms*

11.5 Exchange Rate Forecasting Modelling

Forecasting the exchange rates is a challenging and important task for the modern day traders, people associated with the financial markets and general population across the globe. In this report we will be utilizing the time series concepts to do an analysis and predict the daily exchange rates of the Indian Rupee (INR) against the United States Dollar (USD). We have investigated and compared different techniques like using EMD, ACF and PACF, Neural networks, Support Vector Regression and Additive Regression. We have used the daily exchange rates of the Indian Rupee (INR) against the United States Dollar (USD). The foreign exchange market is the largest financial market in the world and forecasting exchange rates are not solely an important task for investors, but also for policy makers. Since market participant do not have access to future information, they try to model the exchange rate by past information. Forecasting foreign exchange rate is one work that supports to foreign exchange rate risk of commercial joint stock banks. This paper explores the behaviour of daily exchange rates of the Indian Rupee (INR) against the United States Dollar (USD) and forecasts the forex rates for next 365 days. We use statistical models ARIMA, and Neural Network to examine the performance and in forecasting the currencies traded in Indian foreign exchange markets. Daily RBI reference exchange rates from January 1960 – May 2019 were used for the analysis.

***Keywords**: Time series, Machine Learning, Linear regression, Classification & Regression Trees, Random Forest, Support Vector Machines, K-FOLD cross validation, KNN, Neural Networks*

11.6 Customer Demographic Clustering and Pricing for Health Insurance Plans

An existing life insurance company is launching a new set of health insurance plans in India and would like to do a pre- launch analysis on the market base and identify the optimal price range for different types of customers to better aim the marketing pitch at the target audience with attractive pricing option at the first glance. This is not an easy feat considering how insurance products can have varying range of premium amounts and different set of benefits specific to the requirements of customers. Therefore, a sample customer base via survey was collected and the data was used to cluster the market into varying segments and zones to facilitate the marketing and quick pricing. The survey data contains attributes related to the customer demographics, their illnesses or family illness history, physical disability and information on existing insurance plans they may hold. The metrics are then cleaned up and clustered using both k-modes and k-means, to facilitate comparison and it is observed that k-modes returns better results in case of a purely categorical dataset. The survey data yielded 8 distinct clusters and it was observed that almost all the risky customers were picked out to belong in a single cluster which will be of utmost business importance. The model can be expanded to apply across a large customer base as well, and to provide quick premium quotations or to design marketing campaigns.

***Keywords**: Clustering, K-Modes Clustering, Customer Demographics Clustering, clustering in Python, K-modes in python, Clustering in insurance*

11.7 Prediction of Thyroid Condition

This project involves the Prediction of the Thyroid Condition in the Domain of Life sciences. Life sciences discoveries are helpful in improving the quality and standard of life, and have applications in health, agriculture, medicine, and the pharmaceutical and food science industries. This analysis is conducted by using machine learning techniques like CART and logistic regression classification algorithms to predict the condition of thyroid in patients. The project also provides insights on Thyroid diagnosis and detection depending on the given variables. In this project, we are focusing on the prediction of Thyroid condition on potential Humans so that we can help in preventing this from occurring.

***Keywords**: Prediction, Machine Learning, Classification and Regression Trees, R, Logistic regression, Recursive partitioning (Tree based) models – CART, Hypothyroidism and Hyperthyroidism*

12.1 Predict World Cup 2019 Winner by Calculating Team Strength

The primary objective of this project is to explore the possibility of developing a Predictive Model to effectively predict the outcome of a match played in World Cup 2019. We aim to build a model that will provide the below statistics:

- Player Performance Score (Batting & Bowling) - Identify the batting and bowling metrics that provide insight into the performance of a Player – when evaluated as a Batsman and a Bowler
- Team Performance Score (Batting & Bowling) - Derive a methodology to calculate 'Team Performance Score' (separate scores for Batting & Bowling) based on the individual Player's scores.
- Team Performance Score - Identify a method to arrive at a single 'Performance Score' for every team that will take into consideration - the above batting & bowling performance scores, Fielding efficiency & ICC Rankings
- Winning-ness Probability – Predict the chances of winning-ness for both teams in a match.
- Predict the final winner of the tournament

***Keywords**: Predictive Model, Metrics, Probability*

12.2 Drug Store Analytics

One of the major impediments of successfully running a drug store comes from insufficient stocking of medicines based on the demand in the store. In this study we have tried to explore the maximum number of drugs available in the market in consideration to its targeted and measured effectiveness to an array of ailments. Our problem statement encircles the issue faced by store management to have all the medicines available in their store in their limited space. Along with this we have tried to focus on how the store should not lose customers because of unavailability of medicines. The chance that a customer visits the store again will increase if the store can sell all the asked medicines in their visit. In order to start a drug store and run it successfully, the store needs to be started with sufficient medicines required to run the store. However, major problem the store management will face is, it can't have all the medicines available in market inside the store due to limited space. At the same time the store should not lose customers just because of unavailability of medicines asked. So, the drug store needs to keep the most asked drugs with them. The chance that a customer visits the store again will increase if the store can sell all the asked medicines in their visit. In case of chain pharmacies, they have other branches where already business is going on. So, they can use that database for analysis to understand which medicines are frequently sold. In case of hospital pharmacies, the doctors can tell what drugs need to be maintained at store as they are the ones who will prescribe medicines to the patients who visit their hospital. However, in case of independent stores they don't have any database which is already present or any doctor who is dedicated to the store. So, identifying the frequently bought medicines is a bit of challenge in this case. So, in order to identify the most asked medicines for an independent pharmacy we have downloaded data that contains online reviews given by patients who used the medicines. More the number of reviews, more the medicine is used. Better review ratings indicate that the drug is effective and did not cause any bad effects on users. The data is studied and a list of medicines which are used more frequently has been created. Excel, Tableau and R are used for cleaning, exploring the data and building model. The store managers should keep these medicines in the created list at store in order to avoid stock out scenarios and provide customers the medicines asked by them. This improves the customer satisfaction and results in repeat business and loyalty as the customer will be happy if they can get all medicines in one place without roaming around different shops for some of the medicines in the prescribed list.

***Keywords**: Insufficient Stocking, Store Management, Chain Pharmacies, Model Building*

12.3 Yelp Data Analysis

Everyday millions of people visit hundreds of restaurants in US and review them based on their experiences at Yelp. Yelp wants to use this information and classify restaurants into good restaurants and bad restaurants and recommend its users the best possible restaurant in their locality. On the other side, it can also offer business recommendation and consultancy as a service to the local restaurants based on the category in which a particular restaurant falls into.
Yelp has business related information like – business id, business name, ratings, demographic information in business json file and review related information in review json file which has information of the review of a particular business, star ratings and other attributes. Using this information our goal is to understand the sentiment of the reviews and get key insights of the reviews and build machine learning models using Rstudio and python which can classify a particular restaurant into good restaurant and bad restaurant based on the review text and the stars that are provided.
Once the restaurants are classified, yelp can use this information in two ways – Come up with Recommendation to its yelp users and help them choose a better restaurant. The other way is to reach out to the restaurants give them yelp's categorization/rating and offer them consultation as service which will help local businesses to bring up their game.

***Keywords**: Restaurants, Classification, Consultancy, Demographic, Machine Learning Models, Rating*

13.1 Brand Logo Classification

As of January 2019, over 95 million photographs are uploaded on Instagram each day. Social media brings with it a plethora of opportunities for marketers in terms of the sheer volume of images that have the possibility of being uploaded to numerous platforms. Without the availability of tools to sift through all these images to detect their logos in the images, brands will be unaware of the cascade of threats and opportunities directed at them each day. Social media websites contain large amounts of un-labelled image data, from which we can extract many useful insights (such as the share of voice of a digital event and negative brand mention). Vigilant companies closely monitor their web presence by keeping track of any mentions of their brand name on Twitter, Facebook, Instagram and other social media sites. However, a large quantum of social media content consists of user- generated images. We understand that globally, people share almost three billion images on various social media feeds, of which almost 80% of images do not have any textual mention to brands. Instead of tweeting the word 'KFC', a user may instead post an Instagram photo containing a KFC merchandise. The company in question, or third-party analyst, could find detecting such an event beneficial. There is no accompanying text mentioning their brand name. The challenge is that their usual text analytic tools used for tracking the brand will be completely stumped. We fathom that the ability to effectively detect such 'visual mentions' will have many commercial applications soon. The need for logo detection technology allows scanning images for logos to get real uses of products by customers, facilitate monitoring the ROI of marketing campaigns, ensure revenue boost, and more. There are few tools like Brandwatch Image Insights, IBM Image Detection, Amazon Rekognition in the market that have identified this gap. These tools have been using techniques such as Tensor Flow (Machine Learning), Deep learning algorithms to help analyse these images. Researchers have developed various methods for decision making using supervised and unsupervised classifiers. Recently, there has been increased attention to ensemble learning –methods that generate many classifiers and aggregate their results. In this project, we explore a small component of image analytics – classification of greyscale brand logos and building a model to identify the accuracy and sensitivity of classification using advanced analytics techniques such as SVM, Random Forest, GBM on R platform.

***Keywords**: Instagram, Social Media, Image Data, Insights, User-Generated Images, ROI, Machine Learning*

13.2 Assisted Demand Planning using Machine Learning for CPG and Retail

The evolution of computing has offered an opportunity to develop an understanding and shape our environment. In last 1-2 years we have seen that the tipping point has reached where technology is scalable to rapidly leverage the advancements in data science. Artificial Intelligence and Machine Learning is disrupting the way the companies do business by reimagining the efficiency, speed and functionality. In Supply Chain, Demand Forecasting and Planning has been the flag bearer in terms of volume of data it manages. And intelligent automation can help the demand planners to boost their Forecast Value Add process with good results. This project aims to demonstrate the value in using Machine Learning techniques in Demand Planning process.

***Keywords**: Demand Planning, Forecasting, Machine Learning*

13.3 Predict Foreclosure of Loan

The project involves predicting propensity of foreclosure based on lead indicators such as demographics, transactional behaviour and performance on all credit lines. Predicting the probable timeline of the foreclosure has also been included. Customer segmentation has further provided insights on the financial and demographic splits between different sets of customers thus providing more insights. Prediction is conducted by using machine learning techniques like Logistic Regression, Random Forest, Linear Regression and k-means clustering. Significant factors causing Foreclosure of a loan account has been identified and analysed using cost benefit analysis.

***Keywords**: Foreclosure, Home Loan, Balance Tenure, Customer Segmentation, Logistic Regression, Linear Regression, Machine Learning*

14.1 Case Outcome Prediction for SEBI Adjudication Orders using AI-ML Techniques

Predicting case outcome using machine learning models is one of the challenging problems in legal domain. Lack of structured data, usage of handwritten texts, slow in adaptability of latest technology are few of the challenges faced by data scientists working on this domain. Given technological advancement, SEBI (Security Exchange Board of India) had started to maintain their cases online effective 2004. Our aim is to build a Legal analytics solution that can assist SEBI and companies to find their chances of winning/ losing the case at higher accuracy. We use Natural Language Processing tools along with Machine Learning techniques to build the solution. The proposed solution gives 95%+ test accuracy with very good precision, recall and F1score.

***Keywords**: Case Outcome, Natural language processing, Machine learning techniques, Deep Learning, Classification*

14.2 Connecting Students and Industry

The primary focus of this project is to help TIEKE understand and categorise students based on their skill and performance, also by analysing the current job market requirements in online platform we bring out the key skill requirements that recruiters are looking for a fresher job profile. Therefore, by understanding primarily both aspects the student capabilities and job market requirement, it would enable TIEKE to bridge the skill requirement by providing right content and training for students. As a part of their solution currently TIEKE is providing career counselling to students in colleges by analysing their current academic performance with career choice and recommend industry orientation and skill gap

development programmes, the online job portal – FirstNaukri, Internshala and Freshersworld were web-scrapped and the data was used to bring out insights in current skill requirements for fresher's job profile. Hence the project objective is to evaluate the student current academic performance along with career interest and recommend them with skill enhancement content to fit into the job market requirement that is currently trending.

***Keywords**: Job Market, Skill Requirement, Academic Performance, Industry Orientation*

14.3 Building a Recommendation System for Tea Trails Nungambakkam

This project is aimed to use the business insights obtained at a Tea chain in Chennai, which despite having good reviews has failed to bring in repeat customers according to the owner of the outlet. The franchisee model has been successful in the north and western parts of India but is failing to generate more than 25 bills on daily basis. The study aims at identify the reason for this low sales volume and also attempts to find out if there is pattern in the products consumed by the customers. This article applies a version of market basket analysis (MBA) /association rule mapping principles to explore menu items assortments, which are defined as the sets of most frequently ordered menu item pairs (Ping-Ho Ting, 2010).
The data collected has 3702 unique transactions covering approx. 9474 items for roughly 6 months. This dataset suggested that guests ordering a given entrée would most likely then choose a particular side dish to go with it. Findings point to significant opportunities for the client to use existing stored data to better understand purchasing decision patterns that can significantly increase revenue per transaction. Challenges to adoption and future research suggestions are offered. (Dolnicarc, 2016)
Along with finding out patterns the study also looks at rationalizing the menu options as the restaurant has close to 250 items in its menu and it would not be prudent to save inventory for all of the items. The study aims at using RFM methodology and eliminating the inventory which have not been sold for a long time hence not adding up to the revenue of the organization. Indeed, Lee and MacGregor (1985) have presented an analysis of search time that suggests that the optimal number of options per menu panel over a wide range of conditions is in the range of only four to eight pages. (Kenneth R. Paap, 1986).
In this study we have found close to 100 items yield a low RFM score hence we have advised the client to get rid of them. The report looks to demonstrate that it is possible to perform automatic sentiment classification in the very noisy domain of customer feedback data. (gamon, 2004).
The study is made on the online reviews available on Zomato, Google and other online mediums in order to identify customer perception towards the brand. We also carried out an SMS Blast to 24890 respondents yielding 360 responses through which we carried out a survey to cross reference the attribute perception of the clients towards the store. Sentiment analysis/Text Analysis of online restaurant reviews has confirmed the proposed underlying structure of online restaurant reviews (Qiwei Gan, 2016).

***Keywords**: Franchisee Model, Market Basket Analysis, Decision Patterns, Sentiment Classification*

15.1 Inventory and Average Closure Time Prediction for Consumer Cases in India

The general awareness among people on consumer rights has increased considerably in last few years. Government also introduced necessary Acts and provisions to ensure that consumers can exercise their rights in case of any fraud or deceit. All consumer cases across India, for the ease of access and transparency, are available in CONFONET (Computerization and Computer Networking of Consumer Fora in the Country) through NIC (www.confonet.nic.in). As per the data available on this forum, we can see as people become more vigilant, an increase in number of consumer cases can be seen in past few years. However, we think that this increase in cases has an impact both on lawyers as well as on consumers as explained in the below statements which are also the problem statements for our study:

1. This increase in cases could lead to an impact on lawyers/law firms with regards to allocation of lawyers. Hence the objective of our study with respect to this problem is to understand if we can identify any pattern on the number of cases pending for judgement on month by month or quarter by quarter basis. And if we can have a model to forecast the number of cases pending the judgement for a month/quarter, lawyers/law firms can plan their resource allocation accordingly.
2. For consumers there is no way to identify the approx. closure time for their case. Hence, the objective of the study with regards to this problem is to understand if we can have a model which can predict the closure time for a consumer case based on the metrics like state, district and sector.

***Keywords**: Consumer Cases, Judgement, Closure Time Prediction*

15.2 Patent Analysis based on innovations effectiveness in Pharmaceutical Industry

The objectives of this project are to devise a model that will help in deriving trends and insights pertaining to the focus industry (Pharma industry), to forecast its patent granted effectiveness and probable future trends of business losses due to rejection of patents and exhaustive study of the industry fluctuations to understand how innovation helped the company become a top player.

***Keywords**: NLP, Patents grants and applications, R&D, Acquisitions, ROCE, Prediction, testing of model accuracy*

15.3 Personalised Promotions based on Customer Purchase Behaviour

A Grocery store has collected transactional data for households before and after rolling out a promotion campaign. They want a model that is more customer-centric with a focus on customer success first and eventually realizing business success as a result of this model. The data used is a public Dunnhumby's dataset, "The Complete Journey. This dataset tracks household spending over two years from a group of 2,500 households who have been identified as frequent shoppers. For anonymity purposes, the grocery store chain has not been named. Tools used to achieve the objective of the project are R and Python to apply datamining, predictive modelling and recommender techniques. Statistical tests, Market Basket Analysis, RFM, cluster analysis and Recommender system were used. Our objective is to implement a product bundling model, Segment customer base into various customer groups and build a recommendation model from customer purchase behaviour to implement personalized promotions.

***Keywords**: Transactional Data, Customer-Centric, Predictive Modelling, Market Basket Analysis*

16.1 Product Affinity Modelling and Sales Pipeline Enhancement Strategies

The purpose of the project is to segment the customers of a leading B2B IT product company through RFM (Recency, Frequency and Monetary) modelling and to do a market basket analysis of its offerings so that the sales teams can increase the sales through cross-selling and up-selling through these models.

***Keywords**: Customer Segmentation, Modelling, Market Basket Analysis*

16.2 Decoding the Winner Formula on 22 Yards for Winner Prediction

Cricket is one the most watched sport now-a-days. Winning in cricket depends on various factors like home ground advantage, performances in the past matches, experience of the players, performance at the specific venue, performance against the specific team and the current form of the team and the player. A lot of research has been done which measures the team 's performance and predicts the

winning percentage. This report briefs about the factors that cricket game depends on and discusses the various methods which are used to predict the winning of a team in ICC World Cup. The study embarks on predicting the outcome of ICC Cricket World Cup using a supervised learning approach. Study shows team strength between the competing teams forms a distinctive feature for predicting the winner. The data consist of team performances between each World cup with 160 variables with details of batting, bowling, fielding performances, at home and away locations. Modelling the team strength boils down to identifying and classifying team's batting, fielding and bowling performances, forming the basics approach. Statistics and recent performance of a team have been used to model Machine learning algorithms used in predicting the outcome of the tournament.

***Keywords**: Cricket, Performance, ICC World Cup, Supervised Learning, Modelling, Machine Learning Algorithms*

16.3 SCM Analysis-Planning of E-Commerce Fulfilment Centre

With the advancement of technology, traditional business is moving in to online business called E-commerce. E-commerce industry is growing at very fast rate. Consumers are attracted more towards E commerce shopping due to various reasons such as high range of products to choose from, transparency in prices, discounts and cash back offers, free product deliveries, after Sales services and return policy. E-Commerce Industry has many challenges to overcome in order to grow and maintain the business. One of the key challenges is maintaining high service levels to customers, as customers expect deliveries as early as possible.
R was used for programming and churning out the data. Analysis is based on the confidential data from the client. Data mining was done using R. Tableau was used to draw some graphs. Among analytical approach & tools used, KNIME was also used for generating model outputs. K-Means & Hierarchical Clustering is employed on the data to have a cluster dendogram. Microsoft Excel and Word is used to perform calculation and collating the conclusions obtained during the project tenure. To draw the inferences from the data, data analytics life cycle was followed starting from data collection, data cleaning, data exploration and data analysis and finally drawing conclusion. Finally, this report also recommends the future works and use cases that can be done on supply chain data.

***Keywords**: Data Churning, Data Mining, Model Outputs, K-Means, Hierarchical Clustering, Supply Chain Data*

16.4 Predicting Paid Conversions for TradeIndia.com

Online businesses are increasingly bearing high costs (Customer Acquisition Costs) and witnessing low yields on customer acquisition and retention. This project is aimed at identifying key features (variables) for one such web-based entity and attempts to construct a robust-positive-prediction model that can be implemented to improve revenue realization and reduce costs - thereby impacting both the top and bottom line.
At the same time, the focus throughout was to identify key (important) variables and list their importance measures so the business management can build easy-to-implement, daily use rubrics for their operations and sales functions.

***Keywords**: Customer Acquisition Costs, Retention, Prediction Model*

17.1 Application of Machine Learning Techniques for Nifty Index Prediction

Stock market prediction has attracted much attention from academia as well as business. Due to the non-linear, volatile and complex nature of the market, it is quite difficult to predict. As the stock markets grow

bigger, more investors pay attention to develop a systematic approach to predict the stock market. Since the stock market is very sensitive to the external information, the performance of previous prediction systems is limited by merely considering the traditional stock data. New forms of collective intelligence have emerged with the rise of the Internet (e.g. Google Trends, Wikipedia, etc.). The changes on these platforms will significantly affect the stock market. In addition, both the financial news sentiment and Global Indices are believed to have impact on the Indian Stock Index NIFTY. In this study, disparate data sources are used to generate a prediction model along with a comparison of different machine learning methods. Besides historical data directly from the stock market, numbers of external data sources are also considered as inputs to the model.
The goal of this study is to develop and evaluate a decision making system that could be used to predict stocks short term movement, trend, and price. We took advantage of the public economic database which allow us to explore the hidden information among these platforms. The prediction models are compared and evaluated using machine learning techniques, such as Random forest, QDA, Logistic regression and boosted tree. Numbers of case studies are performed to evaluate the performance of the prediction system. From the case studies, several results were obtained: (1) the use of Global Indices historical data sources along to improve the prediction performance; (2) the prediction models benefit from the feature selection and dimensional reduction techniques. (3) The prediction performance dominates the related works. Finally, a decision support system is provided to assist investors in making trading decision on NIFTY.

***Keywords**: Predictive modelling, Data mining, Machining learning, Stock market, Time series, R, Tableau, Capital market*

17.2 Predictive Analysis for Abnormality in Wind Turbine

The objective of this project is to build a predictive model which can predict and also identify the trend in values that can lead to an abnormality in wind turbine. There are various kinds of wind turbine abnormality, generator bearing failure, gearbox failure, main bearing failure etc. Generator bearing failure is a major issue and one measure which is in direct correlation to this failure is generator bearing temp. In this project we are trying a build model to predict generator bearing temp as a collective measure of the significant variables and try to identify the varying trend in the values which may serve as a symptom of an upcoming abnormality. The IMR control charts are used to track the trend. Abnormality can be defined as situations in which the turbine is no longer operating in the way it is intended for maximum optimal power generation. To do this predictive analysis the pedagogy starts right from the initial data gathered to the rightful identification of significant variables that are major influencers on the running status of the wind turbine. The model identified can be used for other turbines in the same wind farm and thus the whole farm is capable of predicting the abnormality.

***Keywords**: Predictive Model, Significant Variables, Maximum Optimal Power*

17.3 Predicting Sales and Improving Profitability of Retail Chain

The project deals with analysis of retail transaction data of a retail chain located at four states of the United States of America, that includes the sales, promotion information for multiple products and brands with different categories for the past 156 weeks. The key challenge is to derive decision that would enable the retail chain to identify the customer preferences out from the high volume of data. The objective of the study to predict the sales for the future, address the relation between price and sales. Also, estimate the impact on sales by changing the price gaps between items, promotions, displays and feature. Also, address price cushion on increase in margin to retailer that would increase the sales. At the end, considerable effort should be spent to identify the products that would enhance the sales and margin to the retailer.

***Keywords**: Sales, Promotion, Products, Customer Preferences, Margin*

17.4 Consumer Behaviour on Milk Consumption- A Comparative Study

This project is about Milk brands used across the states of Tamil Nadu. Through segmentation process we segregate the consumers of the Milk Brands based on demographics factors and Behavioural factors. Data capture from the consumers across the state is done through the Co-operative societies. The survey data is collected randomly, not through any targeted forum or to specific set of people. Insights from this project would help the Milk brands to understand their consumer expectation and target their consumer specifically to improve ROI much better. India has always been the largest producer (an estimated 400 million litre per day currently) and consumer of milk in the world. But it remained a boring market largely because the per capita consumption was low, and most of the milk was consumed in its basic, liquid form, or at best as ghee and some butter. Over the past few years, though, a couple of things have changed to make the market vastly more attractive to new players. India is termed as an 'Oyster' in the world dairy business as its revenue growth is exponential, as of 2017 revenue stands in 80,000 crores. With population segment like India, where most of its nutrients depends on Milk & Milk based product, market is wide open for lot of investment from private brands because most of the Milk production in India is from the village co-operative societies.

***Keywords**: Segmentation, Demographics, ROI, Revenue, Investment*