National Conference

on

Machine Learning and Artificial Intelligence

1 February, 2020

Conference Proceedings

Organized by

Lal Bahadur Shastri Institute of Management, Delhi

In association with

Data Centre and Analytics Lab, IIM Bangalore

&

Analytics Society of India

INTELLIGENT INFORMATION MANAGEMENT USING MACHINE LEARNING AND NLP

Rajesh K Singh,
Director (Technology), CGI, Bangalore, India
raje.singh@cgi.com

Bhattacharyya, Arjun
Director (Delivery), CGI, Bengaluru, India
arjun.bhattacharyya@cgi.com

Avinash Chinchwadkar
Sr. Data Scientist, CGI, Bengaluru, India
avinash.chinchwadkar@cgi.com

Srinivasa Rao, Arvind
Sr. Solutions Architect, CGI, Bengaluru, India
arvind.srinivasarao@cgi.com

While the move from paper to digital information over the past decades has greatly improved information access, it has also complicated information preservation (we create 2.5 quintillion bytes of data every day). The continued generation of business-critical, semi-structured data is changing the storage dynamic because no one could have ever imagined the volume, density and complexity of the records of our times (~51% of data is structured, ~27% is unstructured and ~21% is semi-structured). Without the right data retention and disposition strategy, making investments to extract value from this expanding pool of information can quickly lead to growing corporate risk.

A key element is that, organizations must continually go back and reanalyze the same data sets repeatedly and decide on retention or disposal. Knowledge workers waste 50% of time and 75% of total cost in hidden data as they need to continually look for patterns stretching over hours, days, months, and years.

This proposition brings two important challenges - data classification and clearing electronic garbage. Data scientists spend 60% of their time on cleaning and organizing data and organizations today find it difficult to address these challenges because of the software management layers and tools required to meet compliancy mandates.

This calls for a paradigm shift in the culture of organizations. It also means that summoning technology alone is not the answer. Technology combined with necessary ownership is the need of the hour with 57% IT leaders and 52% IT professionals reporting that they don't always know who owns the data. Empowering user fraternity and making content level insights available to take informed decisions are the best ways to streamline and speed-up this humongous task. Clear visualization focused at the targeted audience plays a critical role and makes sure they get the right insights (45% of the users cite "ease-of-use challenges" as the second-biggest barrier). The new-found ownership will encourage users to align their content, associate correct age to them and discard content that is past its timelines. This approach when taken to the next level, called 'gamification', promotes healthy competition among users. Over time, it addresses the root cause by making sure only relevant content is stored in the first place.

This white paper outlines the solution created for a conglomerate by:

- Providing intelligent document classification to enable business users take informed decisions about retention.
- Using machine learning algorithms to predict retention period and content classification

- Enabling large data visualization by presenting information on single page to enhance user experience
- Gamification technique to enforce ownership and competition

Taking a lead in leveraging above latest technology concepts provides the decision makers with a chance to be a strategic partner with business unit leaders. This new environment represents big opportunities for decision makers.

Keywords: *Document Classification, Prediction of Retention period, Natural Language Processing, Topic Modelling, Logistics Regression, Latent Dirichlet Allocation, Bayesian Generalized Linear Models*

Length of Series and Forecasting Accuracy of ARIMA Models Illustration with Crude Petroleum Prices and EUR/USD exchange rates

K.B.Vedamurthy
Dept. of Dairy Business Management
Dairy Science College
Bengaluru, India
vedandri@gmail.com

Surya Subramanyam
Industrial Bank of Kuwait,
Kuwait City

Rajashekharppa M.T.
SAS, Middle East

For long period time series data, the challenge is to choose the best length of series for obtaining accurate forecasts. Monthly crude petroleum price and the exchange rates (EURO/USD) from Feb,1975 to May 2019 was used to decide optimum length of the series (www.investing.com). Both the series were tested for linearity using the R package nonlinearity Test which employs Teraesvirta's neural network and White neural network tests and found to be non-linear. Therefore, the series were split into linear segments using R package, strucchange. This resulted in four and five structural break points for Crude Petroleum and the Currency prices respectively.

ARIMA models were fitted for different lengths of period based on the indicated breakpoints. The series was divided into two parts a test and a train. The train data covered the respective periods being considered. Testing data spanned the period from June 2017 to May 2019. Models were built on train dataset and validated on the test dataset using accuracy measures such as RMSE, MAPE and Janus Quotient. The results showed that use of recent segments gives better results rather than using all segments. The results show that model with last three segments scored better on all the accuracy measures. However, the model with the recent one structural regime captured the turning points better than the other models, but scored low on the accuracy measures as the forecasts were under estimated.

Therefore, the series was tested for the GARCH effect using the residuals of the model. The forecasts of the estimated GARCH model was used to augment the ARIMA forecasts that yielded better forecasts. To reinforce the above findings the procedure was repeated for exchange rates. Model estimated on recent two regimes yielded better ex-ante forecasts evidenced by the measures of accuracy. Since there was no perceptible GARCH effect, it was not incorporated in the forecasts as in the case of crude petroleum series. Thus, using the recent segments of the given long series has resulted in better forecasts compared to the models with earlier segments.

Keywords*: ARIMA, GARCH, Transfer Function, Forecasting, Mean, Variance, Crude Petroleum Prices)*

Compromised Machine Learning Models – New Era of Security Epidemic

Ashima Purohit
Paypal India Pvt. Ltd
ashimaj@gmail.com

Artificial Intelligence (AI) – and Machine Learning (ML), with its ability to predict based on previously seen data, has become an integral part of modern-day digital applications. Machine Learning models are finding an essential place in virtually every industry, be it medicine, finance, entertainment, law enforcement or cybersecurity, name it and machine learning is there.

Since Machine learning is increasingly being used at the core of several critical applications, such as for self-driving cars, drug recommendation systems, high-volume trading algorithms, privacy and security of sensitive data etc.; any adversarial manipulation on model, can lead to devastating results.

Wondering what would it be like to have your machine learning (ML) model come under security attack, yes COMPROMISED? Have you thought through how to check/monitor security attacks on your AI/ML models?

Historically less attention has been paid to the ways in which AI can be used maliciously. ML models, much like any piece of software, are prone to theft and subsequent reverse-engineering. Machine learning is susceptible to adversarial activity, where an attacker can manipulate the input data to deceive the deployed ML model.

This paper aims to describe the potential threats associated with current methods of collecting or building ML systems and elaborate on the techniques to protect these models. The intention is to bridge the gap between machine learning and privacy and security technologies by helping attendees get acquainted with machine learning, the potential threats to privacy, the proposed solutions, and the challenges that lie ahead.

COMPARATIVE ANALYSIS OF SIMILAR PRODUCTS BASED ON CUSTOMER REVIEWS

Gourab Nath
Assistant Professor
Data Science,
Praxis Business School
Bengaluru, Karnataka 560102, India
gourab@praxis.ac.in

Arpit Prakash
Risk Analyst
HSBC
Bengaluru, Karnataka 560102, India
arpitprakash29@gmail.com

Meenakshi Bhandari
Data Science Consultant
PwC
Mumbai, Maharashtra, 400028, India
meenakshibhandari65@gmail.com

Anuroop Pratap
System Analyst, BI&A
Hexaware Technologies
Chennai, Tamil Nadu 603103, India
anuroopprata p19@gmail.com

The internet has changed the way we do shopping. Online shopping has become very popular due to the increase in access to the internet via desktop and mobile devices. With the advent of various e-commerce platforms, the shopping experience of customers has improved significantly because of convenience, pricing and plethora of products. To enhance the shopping experience, online sellers encourage customers to share their feedback for the products purchased in the form of quantitative ratings, text reviews or a combination of both. These reviews help potential customers to get an idea of the products and compare them with similar products. These also help businesses in understanding customers' perception towards their products and brands. With more and more people preferring to buy online, the number of reviews received for a product could be in hundreds or even thousands. It becomes tremendously difficult for a potential customer to go through all these reviews and summarize the opinions of different customers about the product. Further complexity may arise when it comes to a comparison of similar products. Reading and summarizing reviews about a single product is time-consuming and repeating the same exercise for multiple products is tedious. In this research, we aim to suggest a framework that will enable easier comparison of similar products based on the opinions of the reviewers on different product features. The process will use NLP and Machine Learning to 1) mine the product features of the given products from the customers' review, 2) mine the customer opinions on each product features and 3) put these opinions on the comparative scale for easy comparison.
Merchant sites. However, many reviews are so long and only a few sentences contain opinions about some product features.

For a popular product, the number of reviews can be in the hundreds or even in thousands, which is difficult to be read one by one. Therefore, automatic extraction and summarization of opinion for each feature are required. When a user expresses an opinion about a product, he/she states about the product as a whole or about its features one by one. Feature identification in a product is the first step of opinion mining application and opinion words extraction is the second step which is critical to generate a useful summary by classifying polarity of opinion for each feature. Therefore, we must extract opinion for each feature of a product. In this paper, given a set of customer reviews of a particular product, we need to perform the following tasks: (1) identifying product feature that customer is talking about; (2) extracting opinion words or phrases through adjective, adverb, verb and noun and (3) determining the orientation of the opinion words.

We use a process called part-of-speech (POS) tag to identify phrases in the input text that contains adjective or adverb or verb or nouns as opinion phrases. A phrase has a positive sentiment when it has good associations (e.g., “amazing camera quality”) and a negative sentiment when it has bad associations (e.g., “poor battery backup”).

Telecom Analytics: WAN Outage Prediction

Rahul Pant
Lead Data Scientist
Ericsson Global Services India Ltd
Bangalore, India
rahul.pant@ericsson.com

Priya Hiteshbhai Ghetia
Data Scientist
Ericsson Global Services India Ltd
Bangalore, India
ghetiapriya@gmail.com

Abhi Gupta
Data Scientist
Ericsson Global Services India Ltd
Bangalore, India
abhi.gupta@ericsson.com

In today's society high-speed communication networks play an increasingly important role. The responsibility of network management is maintaining 24/7 network availability and reliability. An intrinsic part of network management is detecting and identifying faults in the network. Network Operations Centres – the centralized monitoring and control stations for network – primarily deal with fault management and performance management to maintain WAN - Wide Area Network - efficiency and customer satisfaction. Some of the major challenges with NOC management today are:

- Troubleshooting billions of service alarms
- Handling ~20 million notifications of workflow management by NOC experts.

A fault in one node can lead to cascading faults in other nodes, resulting in a slew of alarms. Machine learning techniques enable us to discover co-occurring patterns in such a stream of alarms, and other events, which helps to quickly identify the root cause in most fault scenarios. This frees up the NOC operations teams, so they can focus on more complex challenges.

Specifically, in our proposed approach we aim to determine during run-time whether a WAN outage will happen after 10 minutes or not. We have applied Machine Intelligence principles - data mining and data science - to discover patterns of behavior from large historical datasets. These behaviors or patterns essentially mean, correlation between alarms and co-occurrence patterns. The data for around 750 WAN sites over a period of 1 year was collected from Nagios, a network monitoring tool. We used features such as alarm type, alarm severity, average resolution time deviation of the alarms, soft and hard patterns. New features were also generated using rules, patterns and time-based binning. We also used clustering and frequent pattern mining techniques to identify the down and not-down patterns. Several techniques were used to deal with the high data imbalance and noise. Eventually the problem in hand was dealt as a binary classification problem and an appropriate classification model was created with the extracted and generated features. One interesting aspect of our approach is that we evaluated it not only as a time-series data, but also considered how to process the largely symbolic or categorical information collected from the WAN and identify latent behaviours from it.

This approach aids domain experts in learning unknown and evolving patterns of behaviour when the environment is multi-technology and multi-vendor. Such correlated and grouped

patterns enable automatic grouping of alarms which sets the stage for automated WAN outage detection, root causing and self-healing.
Using this approach, we can achieve intelligent grouping of alarms with minimal manual involvement; we can reduce or altogether avoid manual interventions by automatically identifying important and missing groupings and we can reduce the overall number of WAN outages.

Keywords: *alarm correlation, frequent pattern mining, WAN outage, fault prediction.*

Forecasting Agricultural Commodity Prices: Appropriateness of Selected Time Series Models

Deepa M.P.M
PhD Scholar, Department of Agricultural Economics,
University of Agricultural Sciences, GKVK, Bengaluru-65
deepapalb7004@gmail.com

K. B. Vedamurthy
Asst. Prof., Dept. of Dairy Economics and Business Management,
Dairy Science College, Hebbal, KVAFSU, Bengaluru

P. S. Srikantha Murthy
Professor, Dept. of Agricultural Economics,
University of Agricultural Sciences, Bengaluru

Behaviour of agricultural commodity prices are difficult to model and forecast in lieu of the non-linearity present in the series. ARIMA models are widely been used since there inception in the sixties and they have stood the test of time. In recent years non-linear models have made considerable inroad into the domain of agricultural commodity price forecasts. Forecasting of these prices is of critical importance in decision making of the stake holders. Forecasting commodity prices is fraught with difficulties associated with data availability and quality.Time series have evolved from conventional decomposition to the more sophisticated models such as ARIMA, ANN etc. Subjected to the availability of the data even multivariate models have been used with a fare degree of success. However, these models are data intensive with regard to exogenous variables. In this paper, an attempt is made to compare the predictive performance of the linear ARIMA vis-à-vis ANN usingTomato and Onion commodities for their prices.
This paper compares the accuracy of ARIMA and ANN models using MAPE, RMSE and Janus Quotient.The results shown that ARIMA model performs better with lower MAPE value of 12.45 percent and 39.28 percent in Onion and tomato respectively resulting in greater accuracy than the ANN model (MAPE of 15. 61 % in onion and 45.15 % in tomato). This could be because that ANNs require longer series to produce better results. Using seven ears daily data, it was found that ARIMA provedsuperior to ANN, but there was a significant improvement in accuracy of the ANN model from 45.15 per cent to 14.16 per cent.

Keywords*: Price forecasting, Artificial Neural Networks, ARIMA, mean absolute percentage error and forecast accuracy.*

PREDICT MEN'S CRICKET WORLD CUP 2019 WINNER BY CALCULATING TEAM STRENGTH

Netali Agrawal
Infosys Technology Ltd.
Hyderabad, India
netali514@gmail.com

Manav Partap Singh
Tata Consultancy Services Ltd.
Bengaluru, India
manavpratap@outlook.com

Arunava Ghosh
Tata Consultancy Services Ltd.
Bengaluru, India
ghosh.arunava97@yahoo.com

The primary objective of this paper is to explore the possibility of developing a Predictive Model to effectively predict the outcome of a match played in Cricket Men's World Cup 2019. The scope of the paper considers stats of players who played from 2015 World Cup (after) and all the One Day Internationals played till date (excluding 2019 World cup) by the participant nations of World Cup 2019. For all the 10 nations, the study considers only the data of those players who are included to play in World Cup 2019.

Keywords: *Data, Modelling, Implementation, Methodology*

Image Dehazing using UNet and VGG16 Networks

Joxy John
Dept.of Electronics and Communication
Rajagiri School of Engineering and Technology
Ernakulam, India
joxyjohn199515@gmail.com

Swapna Davies
Dept.of Electronics and Communication
Rajagiri School of Engineering and Technology
Ernakulam, India
swapnadavies@gmail.com

Humans are able to see objects as light is reflected from them. This reflected light is absorbed or scattered by the air. If the air contains additional particles like fog, mist, the visibility of the object reduces. The original light is replaced with reflected light which results in haze. Dehazing is the process of removing haze. By using the pretrained neural networks, dehazing can be carried out using the neural network without estimating the value of atmospheric light, transmission map. The pre-trained networks UNet and VGG16 are used for performing dehazing and are modified for regression. Frida dataset is used. The dehazing is carried out in both pre-trained network first with a total of 90 images and second with a total of 330 images. The dataset itself is differentiated into five categories such as homogenous fog, heterogenous fog, cloudy homogenous fog, cloudy heterogenous fog and haze-free images. For each category dehazing is performed using both pre-trained network and RMSE and Loss plots are plotted. The PSNR (Peak Signal to Noise Ratio) of each output image is compared. Better dehazing results in high PSNR value with low RMSE and Loss plots.

Keywords: FRIDA Dataset, Regression, U-Net, VGG16.

Cognitive Testing and Validation Framework for Conversational Bots

Reena V Nagrale
Data Scientist
GTS Analytics, IBM India
Pune, India
renagral@in.ibm.com

Pritpal Arora
Sr. Enterprise Architect
GTS Analytics, IBM India
Bangalore, India
pritpal.arora@in.ibm.com

Current AI systems do not provide an end-to-end testing and validation framework that will also provision continuous and dynamic life cycle of accuracy improvement management for Conversational Assistance (CA) systems like chatbots. Existing methods for training corpus are not scalable and have high degree of dependency on manual intervention with SMEs to validate each findings, result-sets, responses and changes derived from the AI systems.
Cognitive testing and validation framework (CTVF) aim to provide end to end testing, Baseline Analysis and Validation of learning and feedback components across multiple criteria at each stage of an Enterprise Content Management specifically in the context of CA systems.
CTVF automates the testing and validation of supervised Machine Learning Model also generating validation data sets. It further provides a means to continuously evolve the models by dynamically measuring the performance across various metrics at multiple stages ensuring the consistency and quality of modules.
In this paper, we present CTVF accuracy management and training lifecycle methods with different metrics used in validation process in reference of CA systems. The paper will describe different metrics like Precision Index, Similarity Index, SME Relevancy Scoring Index and others as applied in the context of validation processing for CA systems.

Keywords: *Supervised Learning Algorithms, Evaluation Metrics, Text Analytics, Conversational Assistance Systems*

COST MINIMIZATION FOR DATA PROCESSING AND STORAGE IN HEALTHCARE

Biswajeet Padhi
Senior Software Engineer
CGI Hyderabad, INDIA
biswajeet.padhi@cgi.com

Satya Prakash Sahu
Senior Software Engineer
CGI Hyderabad, INDIA
satya.p.sahu@cgi.com

In today's world where data is rapidly growing, 90% of the available data is produced in the last three years. It is not possible to process this enormous data with the existing traditional methods so there is a shift towards using Big Data technologies. This shift has leveraged distributed technologies to process Big Data over multiple workstations. In the present times just processing the data is not enough, the speed at which the data processing happens is of utmost importance. Data processing speed is directly proportional to hardware configuration and workstation costs. The scenario in Health care is no different. The Health care industry often fails to strike the right balance between data, speed, and cost. To reclaim this balance, the size of data should be reduced without losing information from the full data volume, similar to image compression. Using Principal Component Analysis and Auto-encoder Neural Network algorithms of Machine learning and Deep learning we can provide a solution to this problem. These algorithms use orthogonal linear transformation, Eigen-decomposition of the covariance matrix and Singular value decomposition. The K principle component is discovered, and the dimension of data is reduced. This reduced data volume contains nearly the same information as that of full volume data. Now we have a much smaller volume which still conveys the same information, due to less volume we directly strike on the space complexity and by requiring lesser processing time, time complexity also went down substantially.

Keywords: *Machine Learning, Dimension Reduction, HealthCare, Auto Encoder, Neural Network, Deep Learning, Time and Space Complexity*

Improving Learning in First Year Mathematics courses in Engineering Education using Learning Analytics

JAYASHREE TATINENI
Professor
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad, INDIA
Jayashree_t@vnrvjiet.in

SHUCHI TIWARI
Assistant Professor
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad, INDIA
shuchi_t@vnrvjiet.in

KRITI OHRI
Assistant Professor
VNR Vignana Jyothi Institute of Engineering and Technology
Hyderabad, INDIA
kriti_o@vnrvjiet.in

Mathematics is crucial for Engineering problem solving, analysis and design. However student performance in these courses and retention is a major concern among the Mathematics Educators. Learning analytics using predictive models have been used to predict student performance, identify at-risk students, and set up intervention schemes in order to help students improve learning.

In this study, learning analytics is used to track and predict the student's performance and identify potential students who are most likely to struggle academically (at risk). Past student data is analyzed to understand the present challenge of improving the learning and performance in Mathematics courses in first year of engineering education. Future performance of current students will be predicted within the semester in order to prevent them from failing mathematics subjects, giving them instant support that they need for the course progression thereby improving retention. The study will include four cohorts (2016 to 2019, i.e. past, present, and future data) of all first-year engineering students. Datasets will incorporate demographic, socio-economic and student academic performance variables.

This paper is organized as follows. Firstly, relevant literature on mathematics performance, support and learning analytics is reviewed. These are critically evaluated to justify the need for this project. Secondly, the project plan is described, and finally, the paper concludes with expected outcomes of the project.

Keywords: *Predictive modeling, Mathematics Education, Student retention*

PROBABILITY OF DEFAULT ESTIMATION USING MACHINE LEARNING

Tarun Jain
Delhi, India
jaint1991@gmail.com

Shivani Bali
Associate Professor
Lal Bahadur Shastri Institute of Management
Delhi, India
shivani@lbsim.ac.in

It is indispensable for the banks to develop probability of default (PD) models to estimate expected credit losses (ECL) as prescribed by International Accounting Standard Broad (IASB), Financial Accounting Standard Board Risk (FASB) and BASEL to measure credit risk. It is a shift of risk estimation from incurred loss prospective to expected loss where timely recognition of default event is essential and a quantitative model is required to capture it.
This paper presents probability of default (PD) models using nonlinear regressions (Logit, Probit and Complimentary) and focuses on machine learning models viz. decision tree, gradient boosting, survival modeling and lifetime machine learning models as a renaissance. Further, Neural networks can be used for model comparison or benchmarking as it is not a white-box model. It is also useful to find non-linearity in logistic regression if neural networks out-perform over the logistic model. Descriptive statistics, univariate analysis, multivariate analysis, and expert judgement will be the part of the analysis.

It has been found that the Logit regression model has better accuracy than other non-linear regression models and can be investigated further for validation purposes. It is also required to compare the same with the decision tree and machine learning models. Benchmarking, Backtesting, model monitoring and other key things can be analyzed further for the complete demonstration of the models but the heart of this research paper is to develop lifetime PD models conditioned by macroeconomic variables. Further, it is important to analyze the best and worst-case PD scenarios to arrive at weighted PD value for different risk buckets or pools. Last part of this research paper is based on the business strategies where significance of statistical models has been highlighted and practical usage captured to take necessary business decision of loan sanctioning and pricing.

Keywords: *Probability of Default, Regression, Survival Model, Decision Tree, Neural Networks*

GEOSPATIAL ANALYSIS USING TELEMETRY

Geethanjali Battula,
Ford Motor Pvt. Ltd.
GTBC, Sholinganallur,
Chennai - 96

Vibhor Rakesh
Ford Motor Pvt. Ltd.
GTBC, Sholinganallur,
Chennai – 96

Geospatial data is the data referenced to a place – a set of geographical coordinates or simply, data with location information. The location may be static in the short-term (e.g. location of charging station, car dealers, point of interest etc) or dynamic such as a moving vehicle, pedestrian etc. It is an increasingly important source of value creation since this data can be utilized to solve many business problems for customers and Automotive companies. Figure 1 shows a modern day Connected Car which are already "data centers on wheels". Telematics data collected using Embedded Modems or Plug in devices is transmitted through Control Area Network (CAN) Translator to Cloud and then stored in a scalable computing platform for usage. A connected vehicle can generate large Gigabytes of such data per hour of operation [1] and can have up to 100 in-built on-board sensors permanently monitoring speed, engine temperature, braking, spatial data along with variety of other vehicle.

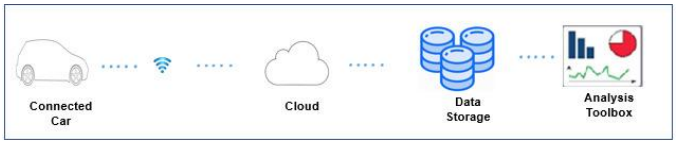

Figure 1: A Connected Car can communicate with the cloud and/or the transport infrastructure and transmit data to the cloud and can be used for various analysis.

This paper presents our research work on how we leveraged this rich geo spatial data collected from Connected Vehicles in combination with other data sets and used geo-spatial techniques which offered great value for customers and enterprise. In the manuscript we highlight a few UseCase(s) such as "Recommending Nearest Charging Station" or "Suggesting Nearby Dealers" which utilize the concepts of geospatial techniques like Geo Fencing, Geohash, Spatial KNN. In addition, the telematics data can provide vital feedbacks for cities and states about traffic volume and roadway design.

Keywords: *Geospatial, location, telematics, services.*

OPTIMIZED PATH RECOMMENDATION FOR NON-CONVERTING VISITORS

Devanathan G
Delivery manager
Nabler
Bangalore, Karnataka
devanathan[dot]g[at]nabler[dot]com

Sushrut Tendulkar
Sr Consultant
Nabler
Bangalore, Karnataka
sushrut[dot]t[at]nabler[dot]com

With the focus to increase the conversion rate of the website, this paper aims to solve two main problems, a) 90% of visitors coming to the website are not converting. How to enable those users to convert? b) Rest 10% of the visitors, who are converting, generally take more than 4 sessions to convert: How to reduce the number of sessions that the converters usually take to convert.

Leveraging the visitor-level data, two segments of the user base, at the visitor level have been considered and for each of the user bases, the multi-session analysis was conducted. We started with studying the behavior exhibited by the converting users at each session and applied those findings to the non-converters. In the second step, we applied the strategies to reduce the number of sessions taken by the visitors to convert, by studying the behavior of look-alike converters. We then analyzed the navigation path of the above-identified segments and looked at where the visitors are falling out of the conversion path in visit 3 and 4. This behavior was compared with those who are converting at visit 3 & visit 4.

Based on the above approach, we identified user characteristics and we ran A/B test for leads visiting back in visit number 2 and 3, provided a personalized direct link to the conversion page from the home page and another link to the product finder page and the product page

Reinforcement Learning for Portfolio Rebalancing

Raamanathan Gururajan
Raam_G@yahoo.com

This paper focuses on building a deep reinforcement learning based agent for efficient portfolio rebalancing in global financial markets. We build a policy based agent to make informed decisions to rebalance portfolios for efficiently managing risk-reward scenarios. We benchmark our results against minimum variance portfolios & equal weighted portfolios, capitalization & price weighted indexes.
We observe our agent to perform competitive against these benchmarks. We conclude by understanding various learnings of our agent, its ability to manage down-side risks and explore the rationale behind its performance.

Keywords: *Deep Reinforcement Learning, Q-Learning, Policy Learning, DQN, RNN, GRU, Actor-Critic, Portfolio optimization.*

Understanding Social Media Users: A Segmentation Analysis

Shubhangi Jore
School of Business Management
SVKM's NMIMS Deemed-to-be University
Indore, India
shubhangi.jore@nmims.edu

Kritika Mathur
School of Business Management
SVKM's NMIMS Deemed-to-be University
Indore, India
kritika.mathur17@nmims.edu.in

Technology was introduced to provide ease to the humans, to help them grow, for entertainment or work but excessive use of everything is also dangerous. Social Media is one of those; normally, it was introduced for entertainment but excessive use of it affects an individual emotionally, mentally or physically. Present study has been conducted to identify the segments of users based on factors of usage of Social Media using Supervised and Unsupervised Learning Algorithm. The results show that around 56 percent of the respondents had a balanced emotion while using Social Media, around 26 percent of the respondents are a group of people who get upset on social media shut downs and feels proud on using the social media and rest all other respondents are sensitive in terms of concerned by. Further, a discriminant analysis was employed to examine whether the groups of clusters are appropriately placed. The ANOVA was used to differentiate the three clusters. It can finally be concluded that there are different perspectives of Social Media usage for an individual, with varied emotional expression. Present study attempts to provide useful implication for marketers and several companies to bring out effective Social Networking marketing program.

Keywords: *Social Media, emotional expression, unsupervised learning, cluster analysis, discriminant analysis*

Identify Relationship between the Products that people buy to increase the cross selling using Market Basket Analysis

Manisha Verma
Assistant Professor (CSE),
Doon Business School, Dehradun, UK, India
mankiishaverma@gmail.com

Market Basket Analysis is the data mining technique used to know the correlation between one item to another purchased. It is used to identify the best possible combination of Products that people buy for example, customers that buy a Bread and Butter are likely to buy Jam or egg. All of these items have been bought in a single transaction. Transactions are analysed to identify rules of association. For example, one rule could be: {Bread, Butter} => {Milk}. This means that if a customer has a transaction that contains a Bread and Butter, then they are likely to be interested in also buying Milk. Before acting on a rule, a retailer needs to know whether there is sufficient evidence to suggest that it will result in a beneficial outcome, we therefore measure the strength of a rule by calculating the following three metrics (Support, Confidence, Lift). This association rule mining technique will help shop owners to identify customer purchasing behaviour and to maximize profit.

Keywords: Market Basket Analysis, Association Rule, Data Mining

NETWORK INTRUSION DETECTION USING VARIOUS AUTOENCODER METHODOLOGIES

Dr. Pattabiraman
VIT University, Chennai
pattabiraman.v@vit.ac.in

Mukkesh Ganesh
VIT University, Chennai
g.mukkesh2017@vitstudent..ac.in

Akshay Kumar
VIT University, Chennai
akshaykumar.satheesh2017@vitstudent.ac.in

Network security is one of the most critical fields of computer science. With the advent of IoT technologies and peer-to-peer networks, the significance of mitigating security threats has never been higher. Network Intrusion Detection Systems are used to monitor the traffic in a network to detect any malicious or anomalous behavior. Anomalous behaviour includes different types of attacks such as Denial of Service (DoS), Probe, Remote-to-Local and User-to-Root. If an attack/anomaly is detected, custom alerts can be sent to the desired personals. In this paper, we will be exploring the effectiveness of various types of Autoencoders in detecting network intrusions. Artificial Neural Networks can parse through vast amounts of data to detect various types of anomalies and classify them accordingly. An autoencoder is a type of artificial neural network which can learn both linear and non-linear representations of the data, and use the learned representations to reconstruct the original data. These hidden representations are different from the ones attained by Principal Component Analysis due to the presence of non-linear activation functions in the network. Reconstruction error (the measure of difference between the original input and the reconstructed input) is generally used to detect anomalies if the autoencoder is trained on normal network data. Here, we used 4 different autoencoders on the NLS-KDD dataset to detect attacks in the network. With just reconstruction error, we were able to achieve a highest accuracy of 89.34% by using a Sparse Deep Denoising Autoencoder.

Keywords: *Network Intrusion Detection; Artificial Neural Network; Autoencoders; Anomaly Detection*

WEBSITE CLASSIFICATION AND PREDICTION BASED ON BROWSER HISTORY

Prakhar
Student
Lal Bahadur Shastri Institute of Management
Delhi, India
Prakhar_@lbsim.ac.in

Rakshit Bhatnagar
Student
Lal Bahadur Shastri Institute of Management
Delhi, India
Rakshit_@lbsim.ac.in

The explosive growth in the amount of available digital information and the number of visitors to the Internet have created a potential challenge of information overload which hinders timely access to items of interest on the Internet. This has increased the demand for recommender systems more than ever before. A web browser should not be only for browsing web pages but also help users to find out their target websites and recommend similar type websites based on their behavior. Throughout this paper, we propose two methods to make a web browser more intelligent about link prediction which works during typing on address-bar and recommendation of websites according to several categories. Our proposed link prediction system is actually frecency prediction which is predicted based on the first visit, last visit and URL counts. But a recommender system is the most challenging as it is needed to classify web URLs according to names without visiting web pages. We add hyperparameter that finds the best parameters for existing URL classification model and gives better accuracy. In this paper, we propose a category wise recommendation system using frecency value and the total visit of individual URL category.

Keywords: *prediction, frecency, classification, naïve Bayes, Random Forest*

Transforming Urban life: Particulate matter emission and dispersion analysis

Soumya Ranjan Mohanty
Dual Degree Student
Mining Engineering Department,NIT Rourkela
Rourkela,India
soumyavicky9@gmail.com

Amiya Kumar Samantaray
CEO
Phoenix Robotix Pvt Ltd.
Bhubaneswar, India
amiya@phoenixrobotix.com

Owing to the rapid industrialisation, vehicular emission and various other combustion activities there has been an upshot in the amount of air pollutants such as particulate matter(PMx),being added to the atmosphere.Depending upon the size in micrometers these particles are classified as PM10,PM2.5,PM1.Notoriously these particulate matter(PMx) are a cause of various respiratory diseases,visibility impairment, environmental and material damage.Thus a genuine case crops up to monitor the emission and the dispersal of these particulate matter in order to identify the sources of pollutants, demarcate the pollution hotspots, help to develop a data based formulation of mitigation measures and provide the general public with reliable data on the extent of exposure to these pollutants. Though there is an evident requirement for wide scale monitoring the implementation of the system has not succeeded on a large scale especially in the developing economies owing to huge installation and maintenance cost, complicated installation procedure of the pre-existing atmospheric parameters monitoring systems. For this purpose the use of cost effective monitoring systems, IoT based network development and end user data availability is gaining in importance. This study analyses various aspects of a cost effective particulate matter monitoring system for wide scale implementation in cities and adjoining industrial hubs to monitor the ambient air quality. It involves the deployment of an IOT based sensor model with a view to develop a network of monitoring devices that can provide highly accurate and reliable field data on particulate matter. It investigates calibration techniques such as neural network models to improve the quality of data acquisition, post processing and finally giving an output using techniques of geo spatial analysis such as krigging. The aim would be to provide reliable data, capture the trends in the change in concentration of particulate matter emission, validate the models, reduce the inter model variability and finally provide the data to end user through data visualisation. This data is to be used by government as well as general public. This will form the basis of formulation of pollution guidelines, monitoring the actions of polluters and provide a way out for the general public to tweak their routine chores to have minimum adverse health impact.

Keywords: *Cost effective sensors, Particulate matter, Calibration models, Neural networks, Geospatial analysis*

Stock Price Prediction Using Convolutional Neural Networks on a Multivariate Time Series

Sidra Mehtab
School of Computing and Analytics
NSHM Knowledge Campus Kolkata, INDIA
smehtab@acm.org

Jaydip Sen
School of Computing and Analytics
NSHM Knowledge Campus Kolkata, INDIA
jaydip.sen@acm.org

Prediction of future movement of stock prices has been a subject matter of many research work. On one hand, we have proponents of the Efficient Market Hypothesis who claim that stock prices cannot be predicted, on the other hand, there are propositions illustrating that, if appropriately modelled, stock prices can be predicted with a high level of accuracy. There is also a gamut of literature on technical analysis of stock prices where the objective is to identify patterns in stock price movements and profit from it. In this work, we propose a hybrid approach for stock price prediction using machine learning and deep learning-based methods. We select the NIFTY 50 index values of the National Stock Exchange (NSE) of India, over a period of four years: 2015 – 2018. Based on the NIFTY data during 2015 – 2018, we build various predictive models using machine learning approaches, and then use those models to predict the "Close" value of NIFTY 50 for the year 2019, with a forecast horizon of one week, i.e., five days. For predicting the NIFTY index movement patterns, we use a number of classification methods, while for forecasting the actual "Close" values of NIFTY index, various regression models are built. We, then, augment our predictive power of the models by building a deep learning-based regression model using Convolutional Neural Network (CNN) with a walk-forward validation. The CNN model is fine-tuned for its parameters so that the validation loss stabilizes with increasing number of iterations, and the training and validation accuracies converge. We exploit the power of CNN in forecasting the future NIFTY index values using three approaches which differ in number of variables used in forecasting, number of sub-models used in the overall models and, size of the input data for training the models. Extensive results are presented on various metrics for all classification and regression models. The results clearly indicate that CNN-based multivariate forecasting model is the most effective and accurate in predicting the movement of NIFTY index values with a weekly forecast horizon.

Keywords*: Stock Price Prediction, Classification, Regression, Convolutional Neural Network, Multivariate Time Series.*

A Predictive Analysis of the Indian Oil and Gas Sector Using Time Series Decomposition-Based Approach

Manjari Mukherjee
School of Computing and Analytics
NSHM Knowledge Campus
Kolkata, INDIA
manjarimukherjee.18@nshm.edu.in

Ashmita Paul
School of Computing and Analytics
NSHM Knowledge Campus
Kolkata, INDIA
ashmita.paul18@nshm.edu.in

Ipsita Bhattacharya
School of Computing and Analytics
NSHM Knowledge Campus
Kolkata, INDIA
ipsitabhattacharya.18@nshm.edu.in

Forecasting of stock prices using time series analysis presents a very difficult challenge to the research community .However over last decade development of many statistical mechanism made analysis of high volume time series data easy . Stock price movements being random in nature, they can be forecasted accurately using robust predictive models. This paper has presented a highly valuable and accurate forecasting framework for predicting the time series index values of the oil and gas sector in INDIA. A time series decomposition approach is applied to understand the behaviour of the OIL AND GAS sector data for the period January 2014-December 2019. On the basis of the structural analysis of the time series, six methods of forecast are designed .These models are implemented to predict the time series index values for the months of 2019.Extensive results have been provided on the performance of each forecasting methods.

Keywords: *Time Series, Decomposition, Trend, Seasonal, Random, Holt Winters Forecasting, Auto Regression (AR), Moving Average (MA), Auto Regressive Integrated Moving Average (ARIMA), Partial Auto Correlation Function (PACF), Auto Correlation Function (ACF).*

Application of Deep Learning Techniques for Precise Stock Market Prediction

Saikat Mondal
School of Computing and Analysis NSHM Knowledge Campus
Kolkata,India
saikatmondal.18@nshm.edu.in

Abhishek Dutta
School of Computing and Analysis NSHM Knowledge Campus
Kolkata,India
abhishekdutta.18@nshm.edu.in

Piyali Chatterjee
School of Computing and Analysis NSHM Knowledge Campus
Kolkata,India
piyalichatterjee.18@nshm.edu.in

The purview of stock price analysis largely depends on the ability to identify the movement of the stock prices and predict the hidden patterns and trends which the market follows. The sole idea is to gain profit from the investments that we make, therefore the more sure we will be with our predictions, safer will be the outlay. Predictions based on stock prices has been a constant field of research work in the past, however, obtaining the desired level of precision is still an engaging challenge. In this script, we are proposing a combined effort of using efficient machine learning techniques coupled with a deep learning technique (like LSTM) to use them to predict the stock prices with a high level of accuracy. We are considering the daily index values of three different companies namely HDFC bank, Tata Consultancy Services, Cipla which are from different segments of the market – finance, IT and medical science. We are using their daily data of previous 6 years (2013-18) to prepare a training model and implement the results on the test data set to predict the closing values of these National Stock Exchange (NSE) listed companies from January 1 to December 31, 2019. For prediction of the patterns of the price movements we are using efficient classification techniques, and for the actual closing values we are using various techniques of regression. Methods which have been implemented involve logistic regression, SVM (Support Vector Machines), ANN (Artificial Neural Network), Random Forest, ensemble learning techniques (Bagging, Boosting). We also use the deep learning technique of Long Short-Term Memory (LSTM) for the prediction of the closing prices of the stocks and then superimpose the accuracy measures by comparing the LSTM results with the other machine learning models.

Keywords: *stock price analysis, machine learning, deep learning, regression, classification, support vector machines, random forest, artificial neural networks, and long short-term memory.*

Thank You