Methods for Effective Formulation of Business Rules in Title Insurance using Machine Learning Algorithms

K. Neelima
Jawaharlal Nehru Technological University
Anantapur

Dr. S. Vasundra
Jawaharlal Nehru Technological University
Anantapur

***Abstract* – Building a dynamic Business rule Engine for Data transformation has been the need of the Hour in many Companies and in various Domains. Such needs are no exceptions for Title Insurance Industries. Building a Business rule engine (BRE) which transforms the raw data into a Standard form/format has never been a one-time development effort for these industries. The data is generally received from heterogeneous sources and in different formats than expected. This BRE is what the Cleansing and transforming does based on the Business logics. Despite this, the smaller change in the input data format leads to incorrect or ineffective transformation leading to Rework in correcting the incorrectly Transformed Data and causing Latency in the availability of the Latest Data. In this Paper, a Literature Review on various ML & DL Algorithms was studied which can further be applied to the real problem in Building the dynamic Rule engine and predicting an appropriate Business Rule based on the raw Data, that can improve the effective formulation of the best fit Business Rule.**

Keywords—CRISP, DM, DATA MINING ALGORITHMS, TITLE INSURANCE, BUSINESS RULES, RULE ENGINE, PREDICTION, CLASSIFICATION, MACHINE LEARNING, DEEP LEARNING,ARTIFICIAL INTELLIGENCE

I. INTRODUCTION

Title insurance is a type of insurance which deals with any losses caused as a result of the possible defects in the property title under consideration. Business rules are the step by step instructions or statements which imposea type of constraint on a specific aspect of the business. The business rule describes the relationships, characteristics and other salient features of a particular field or set of fields. Business rule is used in everyday businesses, especially to define attributes, entities, constraints and all the relationships among the entities. These rules are usually applied for the explanation of any principle or procedure or policy. The input data can be considered only if the business rules are defined and without these rules the input data is just a record and doesn't have any value. Business rules also helpthe decision maker to make the decision or perform specific action against the input data given.The data mining and deep learning techniques can be used to analyze anddetect the structures and the failure patterns associated with each business rule for the given data set. Automatic learning using machine learning features at different levels of the functionality allows a business rule and associated program of function learn the existing data patterns plus other complex data such as partially or totally unknown characteristics. This can be done by mapping the inputs with the outputs directly with the data using machine learning features and with no human intervention. The machine learning algorithms can assist in understanding the data and provide recommendations to build and automate business rules, thereby providing a proactive prediction mechanism. The next section discusses in detail various research articles and the respective outcomes in selecting appropriate machine learning algorithms which can be applied and adopted for predicting the new business rules, as well as for the effective formulation of the rules to minimize the error rates and improve the learning rate.

II. LITERATURE REVIEW

Lidia Contreras-Ochando etal, have presented how to reduce the background Knowledge primitives and build a Background Knowledge by selecting adomain (or) by ranking the primitives for every example. And, thus only the Domain specific background knowledge and associated functions will be applied for the automatic data Transformation[1].

Wolfgang Kratsch et al, have conducted a structured comparison of DL techniques and two of the classical ML Techniques. From the work, it has been observed that DL justifiably outperformed in comparison with ML when the techniques got evaluated on five log repositories over payload and control flow. Also, DL and ML got examined for critical parameters like event-to activity ratio, variant-to-instance ratio, etc. that proved the metrics, viz. accuracy, and F-score better for DL. Hence, the conclusion that DL is better suited for outcome-based predictive modelling [2].

Abhijit Guha et al, had proposed a Hybrid Model for Anomaly Detection by considering multiple classes as Normal or positive using high dimensional text data in the Title Insurance domain. This Approach has combined Traditional one-class Classification Algorithm called OSVM and a deep learning-based, self-supervised,non-linear dimensionality reduction algorithmnamed autoencoder.This model hasmatched the accuracy of the traditional approaches and has shown significant improvement in both training and inference timing compared to traditional ones[3].

Shiliang Sun et al, have clearly demonstrated the influence of machine learning methods and provided a summary of frequently used optimization methods, such as the Relaxation method, and Non-Convex Optimization methods. Many open problems and challenges were also addressed [4].

Rekha Nagar et al, have published a literature review paper that provides a comprehensive summary of machine learning techniques, algorithms and methodologies, as well as, theapplication of those algorithms and tools used to implement various ML algorithms[5].

Dalton Ndirangua et al, have developed a heterogeneous ensemble model for multiclass classification and outlier detection that combines several strategies and ensemble techniques using AdaBoost, random subspace algorithms and random forest as the base classifier. The classifiers built were combined using the average of probabilities voting rule and evaluated using a 10-fold stratified cross validation[6].

Nur Ahmad Wahida et al, have tested a Proposed Predictive model for predicting the remaining time for completion of any task as a useful proposition. In their approach, the independent Variables are categorized in multi-dimensional Space using the Entity Embedding technique. It augments well to reach out to optimize the algorithm in need, especially when it contemplates to predict the next set of activities in the ordained space, opening multi-task prediction[7].

Ruchi Makani et al, in their paper have discussed the Pros and Cons of Various Machine learning algorithms and highlighted the respective features and Constraintsalong with demonstrating the importance of choosing the right approach [8].

Kwang Leng Goh et al, have illustrated various befitting Machine learning Algorithms for filtering Web Spams. In their method, a numerical score based on the Content features of amessage would identify the message as aspam. Content filters scan for words, Header filters scan the header source, Blocklist filters look for suspicious IP addresses, and Rule-based filters apply customized rules to exclude emails that don't conform; hence, all the above assign scores for what they end up with as search outcomes. Based on the findings, if the score from the above passes a certain threshold, then the email is flagged as spam[9].

Ji Zhang review states that Statistical Methods, distance-based methods, density-basedmethods and clustering-based methods can be used for low dimensional data. This paper extends the literature reviewto high dimensional Data as 1) Project the high dimensional data to lower dimensional data using dimensionality deduction techniques; this becomes the pre-processing work for outlier detection, and 2) Setup a scarcity coefficient from density distributions of projections from the data [10].

Ana Azevedodiscusses the pros and cons of three popular data mining methodologies,such as KDD, SEMMA and CRISP-DM [11].

H. Jair Escalante, in his paper has compared methods like distance based, distance K-based, statistical, kernel based, v-SVM and one class SVM. The Author has found that the Kernel Based method seems to be accurate when compared with the other methods. This was done with the data set for stellar population [12].

Komal Patil, described the importance of the data set and the selection of the most appropriate algorithms and the author also emphasized the need to apply ensemble techniques to deal with the complexity of data. The author provided a comprehensive literature review by analyzing more than 13 articles and clearly classified the algorithm based on the types and algorithms such as Random Forest, and ensemble supervised. This means that the data set needs to be thoroughly analyzed to have a good training data which can represent all possible combinations of domain related information so that the ML algorithms can be used to classify and the testing and evaluation would yield better accuracy and reduce the error rates [13].

Maharshi Modi et al, speaks about increasing the accuracy of the prediction and also making the model a right fitone, using a diverse set of machine learning algorithms and techniques such as Extra Tree, Logistics regression, Naive Bayes and Stochastic Gradient Descent. The author explains in detail the various algorithms and applies them on the dataset of house prices and uses the voting classifier to compare and analyze the outcome of the models.The Authorsfurther use the stacking method, definethe meta-classifier and weak learners, and use stackingto merge different types of models. The Authors measurethe performance using a confusion matrix and voting classifier which performsoft voting to get the output. Hence, when we use ensemble methods and deploy various machine learning algorithms and use a voting classifier, we can obtain a high accuracy rate which can also be measured for various metrics such as Accuracy, Precision, Sensitivity and specificity[14].

Shubham Singh and Monika Nag K elaborateon predicting the real-time prices of real estate based on the property market and the demography of the property. Random forest and Linear Regression algorithms are used. The authors usea 24-feature dataset and 6 predictor variables to predict the prices in real time. The Random forest algorithm uses theAdaboost ensemble method which is robust and enables the reduction of noise in data and further contributes to low bias and low variance.A lot of data insight along with powerful ML algorithms is required to predict the housing and real estate prices. The Adaboost based Random forest method plays a major role to help in deriving rules for the predicted price, based not only on an attribute value but also on the presence or absence of the same [15].

Mayank Chaturvedi et al speak about machine learning applications in theRetail, Hospitality, Education, and Insurance sectors. The paper deals with using Artificial Intelligence and Machine learning in the Insurance sector to reach higher maturity levels using chat bots and image recognition. The authorswriteon how Image recognition can be used to identify damaged automobile parts during the process of filing an insurance claim.

The Deep Learning model can identify the vehicle damages, and dents based on the image recognition technique of the uploaded pictures of the vehicle and predict an estimate based on which the client can decide to proceed with an insurance claim or keep his no claim bonus intact.Insurancecompanies can also perform verifications based on reasons for accident claims made by clients due to traffic issues or weather conditions, as these can be readily cross verified using APIs to weather and traffic sites [16].

Kartick Chandra Mondal et alreport that to stay competitive in business, companies need up to date reliable information to remain stay proactive and ahead of competitors. This can be achieved via an automated ETL process. For Automated ETL, data pre-processing is the key.The authors take us through and explain Machine learning based pre-processing approaches to enhance the data quality enabling an automated ETL processing with minimal human intervention.

The automation of the ETL process happens by the generation maps of three different types such as source extraction map, transformation map and loading map. An ETL tool will use scripts and process these 3 jobs.The paper talks about the use of a Custom Rule Engine and Data Preprocessor. Custom Rule Engine is built on machine learning algorithms to classify the data as structured or unstructured and apply the appropriate rules to load or transform the incoming data.

The Data Preprocessor will enhance the data quality by applying various filtering and transformation processes and making it available for the ML application. Automated data processing is achieved through Database Release Automation. The above approach helps in automating the ETL process and providing the clients with quality real-time data with minimal manual intervention, yielding better machine learning prediction results [17].

NoorhannahBoodhun and Manoj Jayabalan in their paper, discuss the use of supervised learning algorithms for Risk assessment. Predictive analytics can be employed in automating the underwriting process in the Insurance industry. The paper talks about the approach to apply the Principal le Component Analysis(PCA) in zeroing down on the key attributes and implement Multiple Linear Regression, Artificial Neural Network, REPTree and Random Tree classifiers on the PCA applied dataset and predict the applicants' risk level.The REPTree classifier technique uses reduced error pruning to build both classification and regression trees which can further be used for classification of the risk.

The Neural Network gives the algorithm an adaptive learning capability of the model which helps in providing a highly accurate risk assessment prediction model. The paper also does a comparison of the results of correlation-based feature selection (CFS) and principal components analysis (PCA) feature extraction for the MAE and RMSE for MLR,Neural Network,REPTree algorithms etc.,and concludes that ML models when used with CFS deliver more accuracy than PCA.The complex formulas and lengthy process of underwriting can now be done faster and made more accurate with data analytical solutions [18].

Ji Zhang, in his literature, speaks about how the most important step of any ML algorithm, namely, outlier detection can cause bias and variance in predictions. The traditional outlier methods along with complex outlier detection methods used for streaming and multi-dimensional datasetsare also discussed. The paper talks about the classification of outliers-based number of data instances and data types. Point Outliers, Collective outliers as based on data instances.

Vector outliers, Sequence outliers, Trajectory outliers and Graph outliers dependon types of data where the outliers lie.Statistical Outlier Detection Methods using Gaussian distribution models and Regression Models are discussed. The Author also talks about using the Histogram based nonparametric methods of detection. Distance-based Methodis another key method to identify what outliers are also compared with.The author concludes by classifying and evaluating different methods under two heads: can the method detect outliers in high-dimensional data space and can it handle data streams. It is also evident from the experiment that the projected outliers embedded in different subspaces cannot be detected by conventional outlier detection methods [19].

Kaimuru, Dalton, et al talk about the challenges of multi class classification. They had developed a heterogeneous ensemble model. They also applied the SMOTE - a synthetic minority oversampling technique to solve the dataset imbalance problem. Ada Boost and Random forest classifiers were combined and withsoft voting the evaluation of the model was performed using 10-foldstratified cross validation.

Correlation, Information gain, Relief, and Gain ratio filter feature selection algorithms were used for a proposed hybrid ensemble method. Removing point-outliers had a positive impact on classification and the overall weighted ROC classification performance increased on the removal of the point outliers.Other algorithms were outperformed by the robust classifier produced using the Ensemble technique. The authors conclude with the following suggestions.

Remove irrelevant or redundant features that have an effect of reducing the point-outliers, thus leading to improved classifier performance. The Presence of outliers contributes to the degradation of the classifier performance. SMOTE resampling improves rare class detection. Better multiclass classification and outlier detection are thetwo bigger advantages of the Ensemble methods [20].

Biswas, Saroj et al, claimed that Data Mining helps organization by extracting useful information from large datasets. The Artificial Neural Network is preferred due to its high performance. The black box nature of the Artificial Neural Network is one of the issues of deploying it in data mining. The authors, in this paper, proposesa rule extraction algorithm from neural network, using classified and misclassified data. The proposed algorithm modifies the existing algorithm, Rule Extraction by Reverse Engineering (RxREN).

The rules are extracted from the trained neural network.Both classified as well as misclassified data ranges are used on the significant attributes of the respective classes. The authors conclude that the proposed rule extraction algorithm, RxNCM is an effective method for data miningusing ANN as it is able to present predictions in a human understandable form. The algorithm works well with large data sets, mixed mode attribute datasets, and uses a pedagogical approach and extracts highly accurate classification rules from the trained ANN.

A Comparison of the proposed algorithm shows higher accuracy, effective, performance than the RxREN5. The Proposed algorithm can be applied to any classification task, such as diagnosis or predictions [21].

Arshad Ahmad et al, witnessing the improvements made in requirements engineering (RE) processes using ML techniques, conducted a systematic literature review (SLR) based on empirical evidence and classified the ML software requirements specific to Stack Overflow. SLR showed that the data extraction process of the (1) Latent Dirichlet Allocation (LDA) topic modeling is the most widely used ML algorithm and (2) Precision and recall are amongst the most commonly utilized valuation methods for measuring the performance of these ML algorithms.

The authors conclude that the SLR study revealed that while ML algorithms have phenomenal capabilities of identifying the software requirements on SO, there are still various issues that will eventually limit their practical

applications and performances. A close collaboration venture between the requirements of engineering and ML communities and researchers can go a long way in the development of real-world machine learning-based quality systems [22].

Ganjarapalli Manasa et al, suggested that Boolean-based Naïve Bayes Classifier can enhance the reliability in messages and improvethe quality of prediction. It is important to consider the quality parameters and other probability methods for effective prediction when using text-based mining methods [23].

V.Sushma Deepthi and S.Vasundraproposed a securemulti keyword searching technique, which can be considered and included for getting the keywords for finding the right document information and also to find the right matches during the data preprocessing time. There exists a need to secure the tokens and features while extracting the data for further processing [24].

III. SUMMARY OF FINDINGS

Data Conformance is an important aspect of various domains and Title Insurance is no exception.AI and ML have taken over manual tasks in Title Insurance business processes. With increasing data dimensionality and in composite population scenarios, the complexity of detecting anomalies and error rate increases. Automated Data Conformance is the least explored.Deep learning, being the fastest maturing technology, can be combined along with traditional anomaly detectors to facilitate and improve the learning rate, and data classification while reducing the error rate.

Artificial Intelligence can be loosely interpreted to mean incorporating human intelligence to machines. Machine Learning can be loosely interpreted to mean empowering computer systems with the ability to “learn”. DL is the next evolution of machine learning. DL can automatically discover the features to be used for classification, as ML requires these features to be provided manually. Two types of learning, namely,inductive, and deductive learning are considered when dealing with business rules.

Data Processing in batch, and Conventional ETL Process causes Latency & the need for Real time Data Transformation; so, there is a possibility of having multiple data quality issues. Hence, it is important to consider the dynamic business rules engine to avoid more reworks and delay with the manual and error correction process. The efficiency and accuracy of business rules needs to be evaluated against different machine learning algorithms to find the optimality and right formulation of the business rules. It is also important to analyze, compare and select theright neural network activation function and algorithms for dealing with complex data, especially in the title insurance domain or any other related domains which dealwith complex data types.

The below Table1 Summarizes Comparative Study on the most important findings and future directions towards this research context in dynamic data Integration.

TABLE I. COMPARATIVE SUMMARY

Title	Summary/Conclusion	Extension/Projection
[1] Automated Data Transformation with Inductive Programming and Dynamic Background Knowledge	This paper uses deductive learning methods and appropriate domain specific functions to ensure data corrections. Domain specific background knowledge and associated functions were used	Recommend Inductive learning methods for setting up the rules for deductive data correction.
[2] Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction	DL justifiably perform better in comparison with ML when the techniques got evaluated on five log repositories over payload and control flow. Also, DL and ML got examined for critical parameters like event-to activity ratio, variant-to-instance ratio, etc. that proved the metrics viz. accuracy, F-score better for DL. Hence the conclusion that DL is better suited for outcome-based predictive modelling.	Furthering what have been said as scope for futuristic work, it is suggested to statistically approach the data to understand how they are distributed as logs in order to churn out patterns which may prove more purposeful in enabling prediction of the outcome not just their course of events but also on their directions.
[3]Hybrid Approach to Document Anomaly Detection: An Application to Facilitate RPA in Title Insurance	Authors have presented Hybrid Approach for Anomaly Detection by Combining Traditional one-class Classification Algorithm called OSVM and a deep learning-based, self-supervised, non-linear dimensionality reduction algorithm named autoencoder.	Converting the hybrid architecture into a single training deep learning-based model with an objective of concept learning by injecting appropriate cost function. A Generative approach of Training the AE for unseen sample generation.
[4]A Survey of Optimization Methods from a Machine Learning Perspective	Authors have clearly demonstrated the influence of machine learning methods and provided summary of frequently used optimization methods such as Relaxation method, Non-Convex Optimization methods. There are many open problems and challenges were also addressed.	Technology of Transfer Learning can be taken into consideration. The inductive learning methods can be applied to deal with data which is insufficient as well as incomplete.
[5]A literature survey on Machine Learning Algorithms	This paper provides the comprehensive summary of machine learning algorithms, application of those algorithms and tools used to implement the machine learning algorithms.	Some of the ensemble algorithms suggested in the paper can be taken further for evaluating the data quality and prediction models.

IV. CONCLUSION AND NEXT STEPS

Based on the extended literature review, this paper concludes that there exists a need to identify outliers in the given input data, detect anomalies while processing the business rules and properly classify, and cluster them into known, Partially Known and Unknown data. The immediate step in this research is to compare various outlier detection techniques/algorithms by experimental evaluation for Data Quality Conformance. In addition to the comparison of the existing ensemble techniques, it is important to identify, formulate and propose a Flexible Novel Methodology of using Data Mining Methods for effective Data Quality Conformance, Classification of outliers and Business Rule Automation. Real Time Case Studies and experiments need to

be conducted by applying the proposed methodology and compare them with real time statistical Inferences for efficiency improvements in Accuracy, Error Rate and Learning Rate when formulating and automating business rules for title insurance.

REFERENCES

[1] Lidia Contreras-Ochando et al,(2020), "Automated Data Transformation with Inductive Programming and Dynamic Background Knowledge",Joint European Conference on Machine Learning and Knowledge Discovery in Databases

[2] Wolfgang Kratsch,Jonas , Manderscheid, Maximilian, Ro'glinger, Johannes Seyfried,(2020), "Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction",Business& Information Systems Engineering (Springer)

[3] Abhijit Guha DebabrataSamanta,(2020), "Hybrid Approach to Document Anomaly Detection:An Application to Facilitate RPA in Title Insurance",International Journal of Automation and Computing (Springer)

[4] Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao,(2020), "A Survey of Optimization Methods from a Machine Learning Perspective",IEEE Transactions on Cybernetics

[5] Rekha Nagar, YudhvirSingh,(2019), "A literature survey on Machine Learning Algorithms",Journal of Emerging Technologies and Innovative Research

[6] Dalton Ndirangua*, Waweru Mwangib, Lawrence Nderuc,(2019), "A Hybrid Ensemble Method for Multiclass Classification and Outlier Detection",International Journal of Sciences Basic and Applied Research (IJSBAR)

[7] Nur Ahmad Wahida ,Taufik Nur Adib , HyerimBaeb,*, YulimChoib,(2019), "Predictive Business Process Monitoring – Remaining Time Prediction using Deep Neural Network with Entity Embedding",The Fifth Information Systems International Conference

[8] Ruchi Makani &B.V.R.Reddy,(2018), "Taxonomy of Machine Leaning Based Anomaly Detection and its Suitability",International Conference on Computational Intelligence and Data Science (ICCIDS 2018) – Elsevier Procedia Computer Science

[9] Kwang LengGoh,Ashuthosh Kumar Singh,(2015), "Comprehensive Literature Review on machine Learning Structures for Web SpamClassification",Elsevier, Procedia Computer Science 4th Intl Conference on Eco-friendly Computing and Communication Systems)

[10] Ji Zhang,(2013), "Advancements of Outlier Detection – A Survey",ICST Transactions on Scalable Information Systems

[11] Ana Azevedo,(2008), "KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW",IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands

[12] H. Jair Escalante,(2005), "A Comparison of Outlier Detection Algorithms for Machine Learning",Programming and Computer Software

[13] Patil, Komal. (2018). A Survey on Machine Learning Techniques for Insurance Fraud Prediction. HELIX. 8. 4358-4363. 10.29042/2018-4358-4363.

[14] Maharshi Modi, Ayush Sharma, Dr. P. Madhavan,(2020), " Applied Research On House Price Prediction Using Diverse Machine Learning Techniques" - INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH

[15] Shubham Singh and Monika Nag K, (2020), "Land Price Prediction using Machine Learning Algorithm", International Research Journal of Engineering and Technology (IRJET)

[16] Mayank Chaturvedi et al, (2020) "Case Studies of Machine Learning Applications in Retail, Hospitality, Education and Insurance Sectors", International Journal of Engineering Research & Technology (IJERT)

[17] Mondal, Dr-Kartick& Biswas, Neepa& Saha, Swati. (2020). Role of Machine Learning in ETL Automation. 1-6. 10.1145/3369740.3372778.

[18] Boodhun, N., JayabalanM,(2018). Risk prediction in life insurance industry using supervised learning algorithms. Complex Intell. Syst. 4, 145–154 (2018).

[19] Zhang, Ji. (2013). Advancements of Outlier Detection: A Survey. ICST Transactions on Scalable Information Systems. 13. e2. 10.4108/trans.sis.2013.01-03.e2.

[20] Kaimuru, Dalton & Mwangi, Waweru &Nderu, Lawrence. (2019). A Hybrid Ensemble Method for Multiclass Classification and Outlier Detection. International Journal of Sciences: Basic and Applied Research (IJSBAR). 45. 192-213.

[21] Biswas, Saroj & Chakraborty, Manomita&Purkayastha, Biswajit & Roy, Pinki&Thounaojam, Dalton. (2017). Rule Extraction from Training Data Using Neural Network. International Journal of Artificial Intelligence Tools, World Scientific. 26. 10.1142/S0218213017500063.

[22] Arshad Ahmad, Chong Feng, Muzammil Khan, Asif Khan, Ayaz Ullah, Shah Nazir, Adnan Tahir(2020), "A Systematic Literature Review on Using Machine Learning Algorithms for Software Requirements Identification on Stack Overflow", Security and Communication Networks, vol. 2020, Article ID 8830683

[23] Ganjarapalli Manasa DivijaSree,S.Vasundra(2020). "Vector-Based Classification Prediction to Geographical Location",International Journal of Future Generation Communication and Networking, Vol 13, 2020.

[24] V.Sushma Deepthi and S.Vasundra(2019), A SECURE MULTI-KEYWORD SEARCH OVER ENCRYPTED DATA IN MOBILE CLOUD COMPUTING, The International journal of analytical and experimental modal analysis, Vol XI, Issue VIII.