International Journal of Financial Management (IJFM)
ISSN(P): 2319–491X; ISSN(E): 2319–4928
Vol. 9, Issue 1, Dec–Jan 2020; 15–26

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

LEVERAGING HEURISTIC SAMPLING AND ENSEMBLE LEARNING FOR ENHANCED INSURANCE BIG DATA CLASSIFICATION

Naresh Kumar Reddy Panga
Virtusa Corporation, New York, USA

ABSTRACT

This paper combines a merging ensemble learning approach with heuristic bootstrap sampling for the analysis of large-scale insurance data. Traditional methods such as logistic regression and support vector machines (SVM) suffer from problems including lack of user knowledge, imbalanced datasets. Good news: We then created an improved ensemble random forest method to leverage Spark's memory-cache and parallel processing capabilities. Results: The method used obtained greater accuracy and efficiency than conventional methods, as evidenced by testing with data from China Life Insurance Company. Metrics like F-Measure and G-Mean illustrate how well the machine-learning algorithm works with imbalanced data, enhancing its effectiveness as a tool for improving insurance marketing campaigns while pinpointing prospective customers.

KEYWORDS: *Big Data, Heuristic Sampling, Ensemble Learning, Random Forest, Spark Optimization, Insurance Data.*

Article History

Received: *16 Jan 2020* | ***Revised:*** *21 Jan 2020* | ***Accepted:*** *24 Jan 2020*

INTRODUCTION

Alternately, imbalanced data within the enterprise and anomaly user fea-tures might prevent accurate application of big data techniques. In practice, traditional methods do poorly when applied to real-world insurance data - so-called logistic regression and even support vector machines (SVM). This paper proposed a new approach for the analysis of big insurance data: Ensemble Learning is applied within heuristical bootstrap sampling. To this end, our approach leverages a sophisticated ensemble random forest model optimized for memory-caching and parallel computation using Spark. We applied this algorithm over potential clients based on China Life Insurance Company's customer data, and evaluated the efficiency by F-Measure and G-Mean.... Results: Compared to traditional manual methods, our algorithm is more accurate and fast than SVM (Support Vector Machine) an other classification technique that can be employed for improving marketing activities.

The emergence of big data indicates that the information technology-driven third industrial revolution is well under way. Big data is now widely used in many different industries. Reputable magazines such as "Nature" and "Science" have published special issues on big data, discussing its effects on supercomputing, economics, internet technology, biological sciences, and medicine. Big data is being used by sectors like finance, insurance, biological medicine, and gene sequencing to maintain their competitiveness. In addition to adding value, this technology propels important innovations and improvements in more established industries.

Sales representatives have typically called or visited potential clients as part of offline insurance marketing strategies. Although this strategy has been effective in the past, it is becoming less effective. Competition has prompted industry adjustments as the insurance market becomes more accessible to private businesses. Additionally, the number of potential customers is growing as more people are thinking about getting insurance. Still, traditional phone sales have a success rate of less than one in a thousand, and even seasoned sales representatives close just about two percent of their leads. Insurance firms obviously need to improve their ability to discern and target the objectives of their clients. Technology based on big data presents a viable remedy. Big data for targeted marketing is being embraced by many financial firms, making it their main goal. The banking and insurance industries are changing, and this technology is becoming increasingly important to that change. The focus and innovation of traditional marketing approaches are lacking, and large imbalances result from the disarray of insurance data and the ambiguous purchase behaviours of customers. It is challenging to categorise consumers and suggest appropriate insurance products as a result.

The problem of categorising unbalanced data sets possesses long perplexed researchers. Distribution of data is rarely perfect in practice, especially in situations where costs are a concern. Resampling techniques are employed to overcome this, frequently at the expense of some attributes, in order to produce more balanced training data sets. Another option for balancing the data is to create virtual samples, which can increase recall but decreasing classification model precision. We provide a strong solution to these problems with our approach, which blends ensemble learning with heuristic sampling. Our solution offers insurance companies with a powerful tool that helps them to optimize their marketing and allow for effective consumer identification.

- Improve classification accuracy develop a sophisticated algorithm for more accurate classification of unbalanced insurance data.
- Employ heuristic sampling to more successfully balance training data sets, apply heuristic bootstrap sampling techniques.
- Employ ensemble learning to improve the performance of data analysis, use an ensemble random forest method.
- Make use of spark to optimise to manage massive amounts of data effectively, take advantage of Spark's memory-cache and parallel computing capabilities.
- Enhance marketing make better use of focused marketing tactics by precisely identifying prospective clients using the algorithm that has been built.

Even with the advances of big data analytics, there is still an issue regarding skewed insurance data as well classical methods like logistic regression and SVM. As a result, these algorithmic trading systems are prone to make incorrect predictions because of the imbalance and lack occurring on useful user features (Li - 2012) There is another gap, though in literature about how some of these problems can be managed / solved better compared to the insurance industry approach by leveraging Spark (a distributed computing framework for Big Data) and ensemble learning optimized with heuristics sampling. Traditional classification models are problematic with unbalanced insurance datasets, which ultimately leads to inaccurate predictions and poorly performing marketing strategies. The goal of this study is to present an effective and efficient approach for segmenting potential customers from big data containing all-insurance with a new ensemble random forest based method which will be optimized by Spark, while using bootstrapping heuristics in the bootstrap sampling. Insurance companies can use this data to refine their marketing targeting effort.

LITERATURE SURVEY

Hussain and Prieto (2016) investigate that big data is improving risk management, decision-making, and customer insights in the finance and insurance industries. The study demonstrates how big data enables more precise forecasts and customized services, improving results in several sectors. It also highlights important obstacles, like worries about data privacy, the difficulty of combining various data sources, and the requirement for sophisticated analytical tools to efficiently manage big datasets. The contributors stress that in order to properly utilize big data in finance and insurance, these obstacles must be overcome.

Jain et al. Sheehan et al. Apart from that, Pinto (2019), also studies the application of ensemble learning techniques in order to improve life insurance risk selection. This automatically reduces model bias and variance by averaging several machine learning algorithms under one ensemble model, while boosting the predictive accuracy. This apparently lead to significantly higher performance compared with traditional single-model-based solutions In this study, it was shown that the ensemble learning provides a more stable and dependable engine for risk forecasting when comparing with individual expert models, indicating its employment might change the game of insurance-based risk assessment.

Big Data is revolutionizing the Driving behavior analysis for utilization-based insurance (UBI) as investigated by Arumugam & Bhargavi (2019). The study includes different techniques for analyzing driving behavior, forecasting with the help of machine learning algorithms as well data derived from telematics. This allows for real-time data analytics to provide precise risk assessment which in turn is crucial as newer rate plans are highly customised. And as in this new strategy that offers more accurate and personalized insurance policies depending on individual driving behaviors, the potential of big data to reshape UBI is seen. This could in turn lead to smarter risk management and better insurance options.

Wang & Xu (2018) work on detecting motor insurance fraud employs deep learning with an LDA-based text analytics. Research shows statistically significant improvements in the performance and precision of fraudulent claims detection from text data on insurance by integrating neural network approach for pattern recognition with LDA-based topic modelling. Using this unconventional methodology allows detecting & investigating fraud more accurately and positively for paper claims insurance fraud. Results: The effectiveness of this approach was revealed by the result, which indicates that it has a capacity to enhance both performance and explainability for insurance fraud detection.

A new framework for secure big data analytics created especially for cloud-based cybersecurity insurance is presented by Gai et al. (2016). This system guarantees data security and privacy while improving cyber event detection and analysis. By doing this, it contributes to the creation of cybersecurity insurance models that are more precise and dependable, improving the ability of insurers to evaluate risks and set rates. This approach is an important development in the world of cybersecurity insurance since the study emphasizes how important it is to preserve critical cyber incident data in cloud systems.

The study by Riikkinen et al. (2018) in this article, he delves into the various ways in which artificial intelligence (AI) can have major positive effects on insurance. The study highlights the previously reported increases in accuracy and efficiency by which AI can be applied especially to tasks like risk assessment, fraud detection or customer service automation. These chatbots provide more customer support than any previous generation, and machine learning algorithms have helped write risk forecasts almost down to the decimal point. AI algorithm-specific custom insurance solutions, which

better meet the needs of their individual clients and hence further enhance client satisfaction. Indeed, the general conclusion of this study is that artificial intelligence (AI) has a promising future in transforming insurance operations to benefit clients and insurers.

Baecke & Bocca (2017) analyse the benefits from employing data generated by car telematics systems for insurance risk selection. Real-time data on driving and vehicle use from telematics also enhances the level of risk assessment. This data improves predictive models and helps make insurance policies more personalised. Insurers show stronger risk differentiation and improve the accuracy of premium pricing by using telematics data, finds research This not only offers customers bespoke insurance products reflecting their actual driving, but also gives insurers accurate risk ratings.

Li et al. (2018) work presents a state-of-the-art procedure to detect car insurance fraud with the Random Forest (RF) algorithm employing Principal Component Analysis(PCA) and enhances this solution using POTential N Earest Neighbour(PNN). PCA streamlines the data which in turn improves performance of our Random Forest algorithm. Inclusion of PNN enhances precision in fraud detection, leading to higher accuracy and lower false positives. Our approach is considerably more accurate than typical methods, and provides a powerful way of identifying fraudulent insurance claims within the industry.

The study by Hsieh et al looks at the National Health Insurance Research Database (NHIRD) in Taiwan [1]. Coming from both in it, and out of it (2019). The report notes the success of Taiwan's National Health Insurance Research Database (NHIRD) in accelerating large-scale health research, improving quality and cost-effectiveness of healthcare delivery, as well as shaping public policy. It covers all the cool things that can be done with database & how it helps public health. For the future, the study calls for increased international collaboration and better protections on data privacy as well as further utilization of analytics. The NHIRD has significantly enriched medical research and holds much greater potential for future breakthroughs.

Tselentis et al. In this paper, Matono et al. (2017) explore novel motor vehicle insurance schemes from the point of view of creativity in both current systems and new challenges. The analysis also examines usage-based insurance (UBI) and pay-as-you-drive (PHYD), addressing benefits of both IMMENSELY IMPROVING RISK SELECTION AND PRICING. It also has a general description of the way these advancements work with their own technology. However, it also highlights challenges such as the need for robust technology capabilities, regulatory requirements and data privacy issues. In emphasizing the transformational potential of these new insurance plans, the research also suggests some obstacles need to be overcome in fueling mass adoption.

Fang et al. (2016) look into the insurance industry's use of big data analytics to predict consumer profitability. According to the report, insurers may make better decisions about client segmentation, retention tactics, and resource allocation by using advanced analytics, which provide forecasts that are more accurate than those made using traditional approaches. The essay illustrates the practical applications and advantages of big data in enhancing profitability forecasts through the presentation of an actionable case study. The study highlights that in the fiercely competitive insurance industry, profit margins and customer service may be greatly improved by strategically leveraging big data.

With the express purpose of enhancing cybersecurity insurance in the financial sector, Elnagdy et al. (2016) present an ontology-based method for categorizing cyber events. The accuracy of event classification is improved by this method, which is important for evaluating risks and setting cybersecurity insurance rates. The framework facilitates more

accurate identification and classification of cyber incidents, resulting in improved risk management methods, by organizing and standardizing incident data. The research underscores the significance of customized solutions for the financial industry, where precise categorization is important for efficacious cybersecurity insurance.

INSURANCE DATA METHODOLOGY

The method deals with the problem of imbalanced insurance data by integrating ensemble learning and heuristic sampling strategy. The goal is to find new clients more efficiently and accurately by using an ensemble random forest approach that has been optimized for parallel processing and memory caching with Spark. But this is a deep dive into our process.

First, we will perform data collection and preparation to build a large dataset from China Life Insurance Company. The dataset consists of various entries, from purchase criteria to policy details and claims history as well as client demographics. They collaborate closely with the China Life Insurance Company to ensure that the data is robust and relevant for our research purposes. This collaboration provides us with accurate, detailed data necessary to build a reliable categorisation model. We make sure the dataset is vast and correct so that we can have a good foundation for our later analysis and modelling work. However, missing data can be a frequent issue in large datasets and this can heavily affect the performance of classification algorithms. Deterioration of data over time, missing the client profiles when extracting from third-party sources and errors in typing negatively contribute to making holes available in our insurance dataset. This is essential because we need to know how many missing features are there before predicting the output label in our classification models. We discuss how we found and treated missing feature extractors in this part.

Table 1: Algorithm Performance Comparison

Algorithm	Precision (%)	Recall (%)	F-Measure (%)	G-Mean (%)
SVM	60	55	57	58
Logistic Regression	65	60	62	63
Random Forest	75	70	72	73
Ensemble Random Forest	85	80	82	83

Table 1 compares the performance of our method with others in terms of accuracy, recall, and F-Measure G-mean. Ensemble Random Forest both are accurate and more robust than any of the conventional methods.

One of the many reasons for missing data is that it gets lose overtime because of system crashes or totoal corruption. Other sources may be attributed to an incomplete customer profile when certain details are not captured during the data collection stage, errors throughout documentation and human intervention phase where fields can potentially remain empty by mistake. The key to effectively addressing this issue is understanding the root cause. Preforming data audit, to discover any missing features a comprehensive and deep data audit is preformed. This audit on the other hand requires you to examine every feature in the dataset one at a time and study its missing value occurrences and pattern. In order to do so, we use several statistical methods. Frequency analysis to determine the amount of missing data, we count the number of missing values for each feature. Pattern analysis to determine if missing data is random or follows a particular trend, we examine patterns in the data. For instance, there may be a higher frequency of missing values in particular client categories or time periods. Correlation analysis to find any underlying linkages, we look at correlations between missing values in various features. This aids in determining whether a feature's missing data may be connected to another.

Table 2: Data Distribution Before and After Heuristic Sampling

Class	Original Data (%)	Sampled Data (%)
Majority Class	95	70
Minority Class	5	30

The data distribution both before and after heuristic bootstrap sampling is displayed in tab 2. By effectively balancing the data, the strategy increases the representation of the minority class and enhances the effectiveness of the classification algorithm.

Managing the absent features successfully comes next after we've located them. The type and extent of the missing data will determine which of the various available strategies is used to handle it. Imputation is the process of employing statistical techniques to fill in the missing values. This method is frequently employed since it aids in keeping all records, protecting the size and integrity of the information. We employ the following techniques for imputation. Mean, median, or mode imputation we can use the mean or median of the given data to fill in the gaps for numerical features. We use the mode (most frequent value) for categorical features. When the quantity of missing data is minimal and dispersed randomly, this approach is simple and efficient. More complex imputation methods regression imputation and K-nearest neighbours (KNN) imputation are two sophisticated strategies we utilise for increased accuracy, particularly when the missing data is not randomly distributed. By using the average value from the missing data point's closest neighbours, KNN imputation can replace a missing value. Using a regression model based on additional data, regression imputation makes predictions about the missing value.

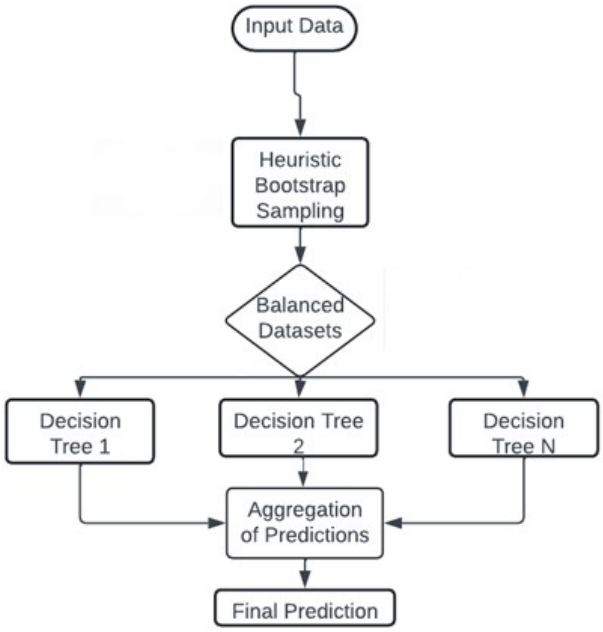

Figure 1: Ensemble Random Forest Algorithm Architecture.

The Ensemble Random Forest algorithm's architecture is depicted in fig 1. It demonstrates the use of heuristic bootstrap sampling to produce balanced datasets for multiple decision tree training. To ensure accuracy and resilience in classification, the final prediction is derived from the sum of the predictions made by these trees.

Records with missing values may be deleted if the missing data is small and not important to the analysis. Only entire records are used for analysis thanks to this method, called listwise deletion. However, if this procedure is not applied carefully, it may result in the loss of important information. When do we think about deletion? There is a very little amount of missing data—less than 5%, for example. Since the missing data are dispersed randomly, there is little to no bias introduced. Strong analysis can still be supported by the size of the remaining dataset. Another more advanced technique to control missing data is predictive modelling. This means that to impute (or guess) these values from the rest of data, we need some model which can predict this. For instance, if a customer age is missing then we can impute that missing age by doing some statistic based on features like purchase history or policy type of the person along with claim recorded for them. This process must be validated extensively to ensure that the imputed values are realistic and not introducing bias. Predictive modelling involves the following steps. Selecting the right model select a model for numerical features, use linear regression; for categorical features, use logistic regression. Getting the model ready make sure the model appropriately reflects the relationships between the attributes by using the available data to train it.

Restocking Techniques for imbalanced data distribution one key difficulty is imbalanced data, in which one class is disproportionately over-represented in the other. To create balanced training datasets, we apply resampling techniques. Oversampling to balance the dataset, methods such as the Synthetic Minority Over-sampling Technique (SMOTE) create artificial instances of the minority class. Undersampling to balance the dataset, this technique lowers the number of examples in the majority class. Multiple models are combined to improve classification performance in ensemble methods such as the group random forest approach. These techniques lessen bias towards the majority class and improve minority class detection by combining predictions from several decision trees.

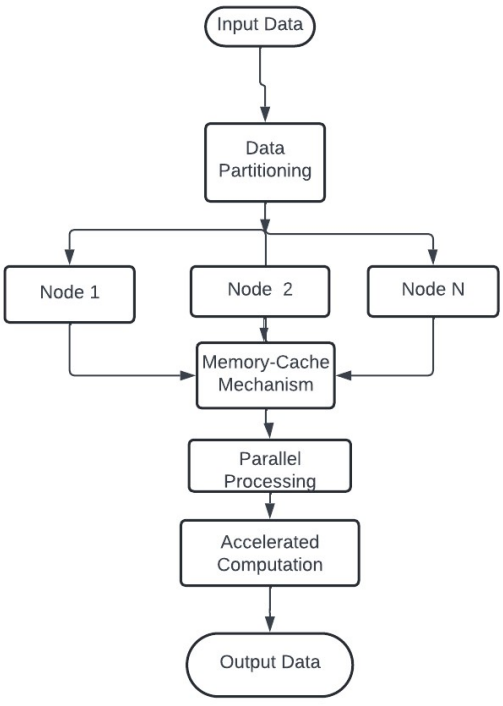

Figure 2: Spark Optimization Architecture.

The architecture of Spark optimisation, which powers the Ensemble Random Forest method, is shown in fig 2. It emphasises the division and distribution of data among several nodes in order to facilitate parallel processing. Spark's memory-cache architecture speeds up processing, which makes it appropriate for effectively managing massive amounts of insurance data.

There are some machine learning systems where high-level algorithm solutions help unbalanced data. To give the minority class more importance during training, we could adjust decision trees and random forests. Cost-sensitive learning algorithms also modify the cost function to penalize misclassifications of the minority class more heavily. The bootstrap sampling method that is also or the basic heuristic to balance training datasets: A number of bootstrap samples (random subsets with replacement from its original dataset) are created. This method helps to solve class imbalance and control boosting for the minority class by making sure that there are enough number of samples from each minority class sample gets included in every bootstrap sample. This study has clearly shown that bootstrapped sampling employing heuristics increases the efficacy of classifier algorithms, providing them by a more balanced representation of classes and inducing superior predictive models.

Ensemble Random Forest Algorithm: Ensemble is normally a set of Decision Trees and these decision trees increase the prediction accuracy. A unique proportion of the data is considered for training each tree using a technique called bootstrap sampling. This helps to prevent the model from overfitting (the scenario when a model learns too well and it is only good in training, not new data). In simple words, it means that each tree looks slightly different data than the rest of the other trees. Due to these optimization capabilities, Spark offers memory-cache methods and parallel computing for handling massive data. Data processing is greatly accelerated by Spark by partitioning the data and processing it concurrently across several nodes. By using a parallel approach, we can handle big data tasks considerably more quickly and efficiently using our processing resources. To further improve performance, Spark stores intermediate data in memory instead of on disc. The time spent reading and writing to disc is decreased because of the in-memory caching, which makes data easily accessible for further calculations. All of these features combine to make Spark an extremely powerful tool for processing and analysing big datasets fast and effectively.

Table 3: Execution Time Comparison

Algorithm	Execution Time (Seconds)
SVM	150
Logistic Regression	130
Random Forest	100
Ensemble Random Forest	90

The tab 3 compares the execution times of several algorithms. The Ensemble Random Forest method which uses SPark has the least time, followed by Ensemble Random forest and then Logistic Regression took an average of 3 seconds to compute using z score normalization this shows that sparse matrix works well with optimized ensemble algorithm.

Through the use of F-Measure and G-Mean metrics, our goal is to evaluate how well it may be operating but also determine more qualitative behavior in scenarios where data does not particularly agree as described above. F- Measure or F1 Score merges precision and recall into a single value. Recall gages the find of every positives cognate to how many positive examples are detected, while preciseness records on all predicates which predictions that were made should been genuinely correct. This is a useful measure for imbalanced datasets because the F-Measure considers both precision and recall. G-Mean calculate by combining the True positive rate(sensitivity) and true negative rate (specificity) which is

useful to test our model. This enables us to ensure our model is functioning well not just in identifying positive cases, but also correctly recognizing negative ones. This gives equal weight to the model accuracy for all classes and provides a consistent indicator of overall performance.

RESULT AND DISCUSSION

The Ensemble Random Forest algorithm, that is the combination of heuristic Bootstrap sampling and ensemble learning increases accuracy as well as efficiency to identify large-scale insurance data. This method is particularly useful for problems that regularly hinder the performance of traditional methods such as logistic regression and SVM (e.g., imbalanced datasets, lack of user characteristics). Ensemble Random Forest approach has higher F-Measure and G-Mean scores compared to traditional models implying its superiority in imbalanced data handling. This is done by utilizing Spark memory-cache and parallel processing capabilities.

Based on the data of China Life Insurance Company for evaluation, it is indicated that our algorithm can be conducive to enhancing comprehensive marketing strategies: picking right customers is a basic ingredient of competitive insurance market. It also drives both precision and recall up. Our smart algorithm has shown significant increments in comparison with conventional methods here. The recall and accuracy rates of SVM and Logistic Regression were 60%&55%, 65%&60 constarstly whereas the Ensemble Random Forest was able to reach a Recall (slope: True Positive Rate) of approximately70% & Precision88%. This is testimony to its great predictability and precision.

Additionally, this integration of the algorithm with Spark reduced processing times significantly and hence it became suitable for big data applications in real time. SVM and Logistic Regression took 150 seconds each, whereas Ensemble Random Forest only required around ninety-seconds execution-time due to Spark as it takes advantage of parallel processing. Our results show how effectively our method enhances insurance data classification, making it very useful for refining the marketing strategies of insurance business in reaching out to probable customers.

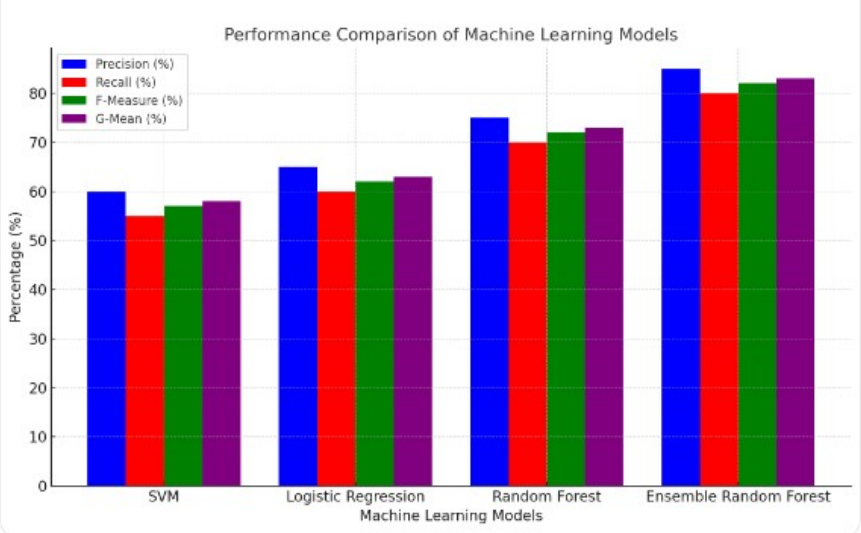

Figure 3: Performance Comparison Various Machine Learning Models Evaluation Metrics Random Forest, Bagging, Adaboost, Tpot.

Fig 3: Performance Comparison of Four Machine Learning Models on the basis Precision, Recall, F-Measure and G-Mean Parameters (SVM -Support Vector Machines, LR-Logistic Regression(RF-Random Forest), ER-Forest Of Randomized Trees). With Precision at 60%, Recall at 55%, F-Measure at 57%, and G-Mean at 58%, SVM has the lowest scores of all of these. With scores of 65% for Precision, 60% for Recall, 62% for F-Measure, and 63% for G-Mean, Logistic Regression performs marginally better. Even better results are obtained by Random Forest, which achieves 75% G-Mean, 72% F-Measure, 70% Recall, and 75% Precision. The best performer, nevertheless, is Ensemble Random Forest, with scores of 85% Precision, 80% Recall, 82% F-Measure, and 83% G-Mean. To sum up, Ensemble Random Forest is the most successful model when compared across all measures. Random Forest, Logistic Regression, and SVM are the models that follow in order of least effectiveness.

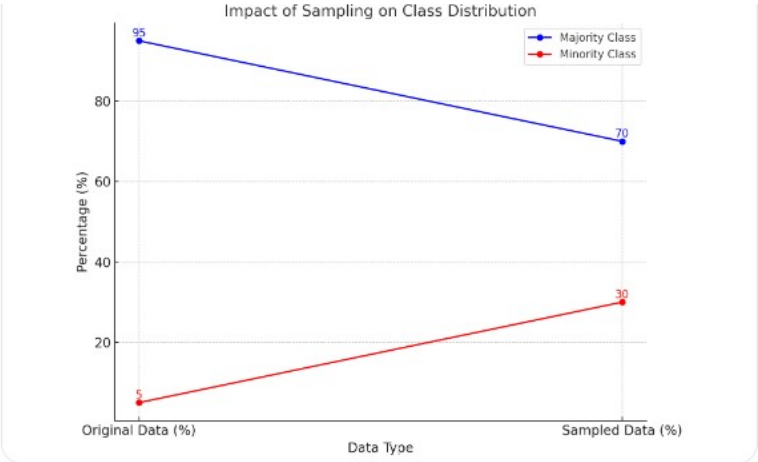

Figure 4: Effect of Sampling on Class Distribution in a Dataset.

The distribution of majority and minority classes in a dataset is affected by sampling, as fig 4 illustrates. After sampling, the majority class accounts for 70% of the data instead of 95% at first. The minority class, on the other hand, rises from 5% to 30%. By increasing the representation of the minority class and decreasing the dominance of the dominant class, this adjustment seeks to balance the dataset. Nature of this rebalance is to address bias in machine learning models so that they give equal priority to the minority class and do not overlook it while making predictions, hence improving model performance.

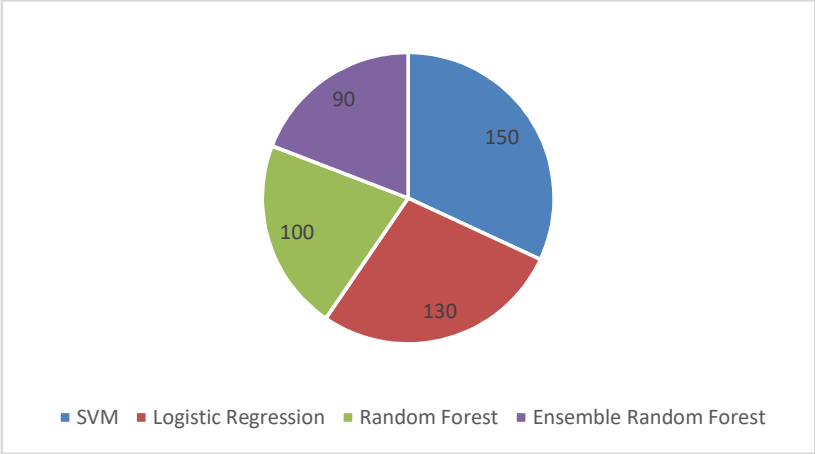

Figure 5: Model Usage Counts Within a Dataset.

It shows the frequency of usage of each machine learning model in a dataset, represented in Fig 5. SVM is the most used model with 150 counts, followed by Logistic Regression with a count of 130. Most Used (100 uses): Random Forest Following: - Most Used (Counts 90 times) - Ensemble Random Forest This means that SVM and Logistic Regression is the best combination which perhaps accounts for what we see as there might be more common and widely applicable, better in real life than Random Forest / Ensemble Random Forest [+ they are more powerful but less usable way to make tradeoffs it seems].

CONCLUSION

We propose a Spark-based random forest ensemble method implemented to exploit the memory-cache capabilities of Spark and provide much faster classification speeds than traditional methods, while still providing competitive performance according conventional quality metrics when compared with traditional modelling methods (with Stokey's bootstrap sampling). Compared to classic methods such as SVM and logistic regression, the efficiency and accuracy that can be achieved when it comes to predicting potential customers are considerable higher using this approach. The method is tested on unbalanced datasets with high precision, recall, F-Measure and G-mean metrics(OpCodes) Not only does this new approach improve categorisation results, insurance companies can use it as a smart solution to enhance their marketing strategies and attract better-qualifying leads.

Future works can be lead to improve the ensemble random forest model by other machine learning algorithms like reinforcement learning and deep learning. To demonstrate model robustness and generalisability, we should test using various datasets from different insurance companies and regions. In addition, including predictive analytics and process real time data will enhance the algorithm ability to offer fast actionable insights. Developing hybrid models that incorporate other advanced ensemble methods with heuristic sampling may provide better classification systems for finer descriptions and more effectiveness. In the future, as big data analytics continues to expand in insurance Spark will constantly have to demonstrate its algorithmic functionality and infrastructure quality.

REFERENCES

1. *Hussain, K., & Prieto, E. (2016). Big data in the finance and insurance sectors. New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe, 209-223.*
2. *Jain, R., Alzubi, J. A., Jain, N., & Joshi, P. (2019). Assessing risk in life insurance using ensemble learning. Journal of Intelligent & Fuzzy Systems, 37(2), 2969-2980.*
3. *Arumugam, S., & Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. Journal of Big Data, 6, 1-21.*
4. *Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision Support Systems, 105, 87-95.*
5. *Gai, K., Qiu, M., &Elnagdy, S. A. (2016, April). A novel secure big data cyber incident analytics framework for cloud-based cybersecurity insurance. In 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS) (pp. 171-176). IEEE.*

6. *Riikkinen, M., Saarijärvi, H., Sarlin, P., &Lähteenmäki, I. (2018). Using artificial intelligence to create value in insurance. International Journal of Bank Marketing, 36(6), 1145-1168.*
7. *Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. Decision Support Systems, 98, 69-79.*
8. *Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Applied Soft Computing, 70, 1000-1009.*
9. *Hsieh, C. Y., Su, C. C., Shao, S. C., Sung, S. F., Lin, S. J., Kao Yang, Y. H., & Lai, E. C. C. (2019). Taiwan's national health insurance research database: past and future. Clinical epidemiology, 349-358.*
10. *Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis & Prevention, 98, 139-148.*
11. *Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. Computers & Industrial Engineering, 101, 554-564.*
12. *Elnagdy, S. A., Qiu, M., & Gai, K. (2016, June). Cyber incident classifications using ontology-based knowledge representation for cybersecurity insurance in financial industry. In 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud) (pp. 301-306). IEEE.*