Data Science in Fraud Detection

Fraud, defined as the crime of getting money by deceiving people is as old as humanity (Cambridge Dictionary). However, modern technology and internet transactions are creating new opportunities for committing fraud, resulting in the loss of billions of dollars worldwide each year. Data science techniques such as machine learning and artificial intelligence can be applied to detect fraudulent activity online. Nuno Carneiro et al. wrote a research paper titled "A data mining-based system for credit-card fraud detection in e-tail" [1], in which they developed and implemented a fraud detection system for a large online retailer. I will be using this research paper as a case study to show how modern data science techniques are being implemented to detect online fraud.

There is a perception among criminals that fraud is a victimless crime and has little impact on society. This perception stems from the belief that victims of fraud are unaffected as they get their money back, with the burden of fraud being passed on to large corporations who can absorb the losses. However, this argument doesn't take into account the devastating phycological impact of fraud on victims. Fraud also has very significant financial impact on small businesses, whose very existence can be threatened by the financial losses [2]. According to a 2014 ACFE report, organizations lose 5% of revenues each year to fraud [3], the cost of this loss of revenue will be directly passed onto consumers. Therefore, we all shoulder the burden of fraud, even if we aren't personally the victims.

The internet provides the basis for the efficient transfer of data between individuals around the world. This allows cybercriminals to commit crimes across international borders whilst in the safety of their own homes. Cybercrime and online fraud are very difficult to police, especially using traditional methods. Cybercriminals use VPNs and proxies to hide their identity and their location [4], meaning that authorities can't identify a suspect for a crime. Even if a cybercriminal was identified, they are often from another country which operates under a different jurisdiction. Therefore, the police can't apprehend the suspect unless they're given permission by the foreign government, which is usually denied. However, cybercrimes do leave breadcrumbs in the form of data. Therefore, analysing data through the use of data science techniques is crucial for identifying and reducing cybercrime such as online fraud.

Identifying fraud can be treated as a classification problem, i.e. identifying if a new transaction is fraudulent or legitimate. To do this accurately, a model must be trained to detect patterns in the dataset, and therefore be able to detect anomalous, and thus fraudulent, transactions. However, fraudulent transactions are difficult to detect because fraudsters try to make their behaviour look legitimate and the number of legitimate records is far greater than the number of fraudulent cases [1]. The performance of different machine learning models in detecting fraud varies according to the situation that they are applied to. Therefore, there is no one machine learning method that is best at detecting fraud, but a range of techniques that should be implemented according the situation. Machine learning techniques are divided into supervised and unsupervised methods. Both of these methods use past observations of transactions in order to train an algorithm. Supervised methods require that each of these observations has a target label, i.e. if its fraudulent or legitimate, whereas unsupervised methods do not. Supervised methods, such as artificial neural networks, support vector machines and logistic regression techniques are popular in detecting fraud in banking and credit card operations [5].

The online retailer in the case study sells items from 400 boutique stores on a commission basis. The company wants to integrate a machine learning model into their fraud detection system in order to raise automation levels to 80%. Fig.1 shows how the proposed fraud detection system of the company will operate. The company requires that the system has less than 1% of accepted orders that are fraudulent and that less than 4.5% of orders are refused.

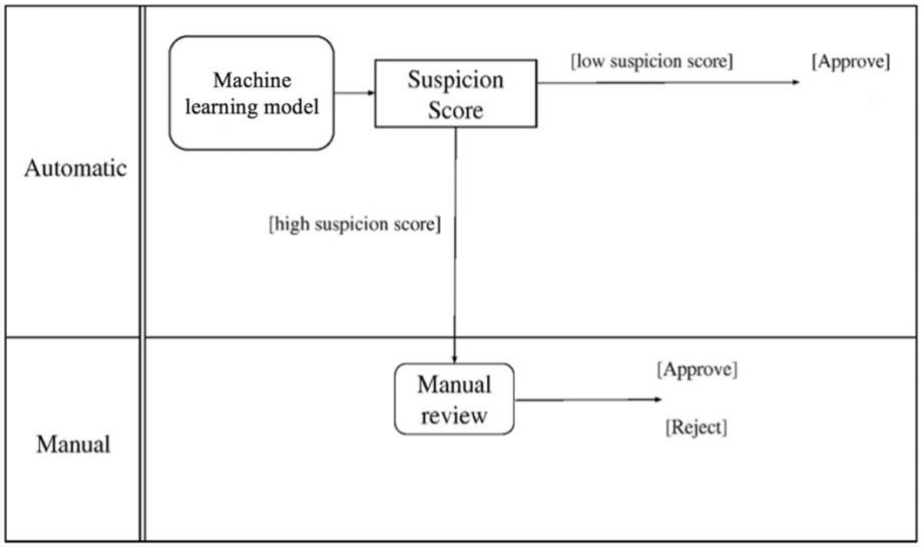

Fig. 1 Proposed process for fraud detection at the company (adapted from the case study [1])

In order to train and test a machine learning model, over 400,000 past orders between January and August in 2015 were prepared and cleaned to form a suitable dataset. Creating and cleaning a dataset of this size is not a trivial activity and involves decisions which can greatly affect the quality of the project [1]. Many data preparation tasks were performed including the transformation of categorical variables into numerical values and standardising the dataset using the Min-Max technique. Each observation in the database has 71 descriptive variables, which cover all of the data gathered in the processing of the transaction, and a target variable determining if the order was fraudulent or legitimate. In order to train a model and evaluate its performance, the dataset was split into a training set and a test set at an 80:20 split ratio.

The random forest model was chosen to be implemented as analysis showed it to be the most effective supervised machine learning technique for this problem. Fig. 3 shows how a random forest model is formed. The parameters of a decision tree such as Min. split are varied, and the different suspicion scores are averaged. As shown in Fig. 2, a decision tree splits a dataset into subsets based on their characteristics in such a way that the entropy, a measure of uncertainty in classification, of each subset in minimised. The output of a decision tree is a suspicion score ranging from 0 to 1.

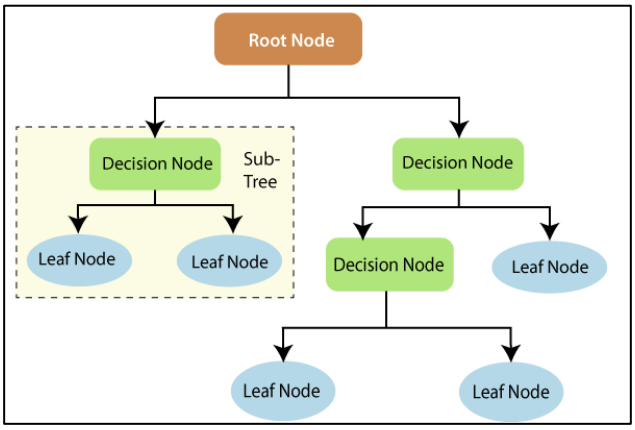

Fig. 2 A decision tree model (From Tony Yiu, "Understanding random Forest") [6]

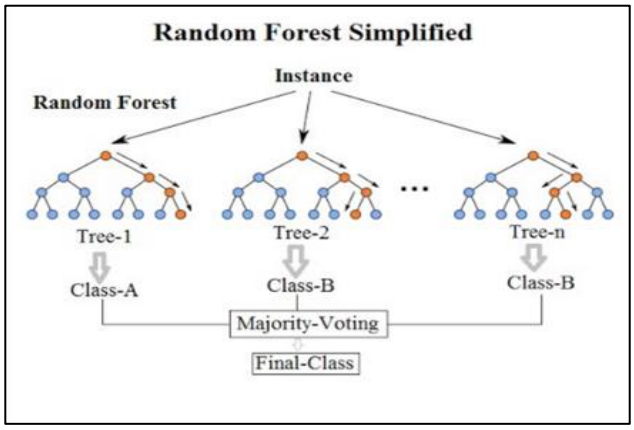

Fig. 3 A Random forest model (From Venkata Jagganath, "Random Forest Template For TIBSCO SpotFire")

The performance of the model was analysed on the test set. To do this, a threshold level on the suspicion score must be defined in order to determine the final classification of an observation. The output of a classification model falls into one of four scenarios: a true positive, a true negative, a false positive and a false negative. Fig. 3 show how the relationships between key performance indicators vary according to the threshold level set on the suspicion score.

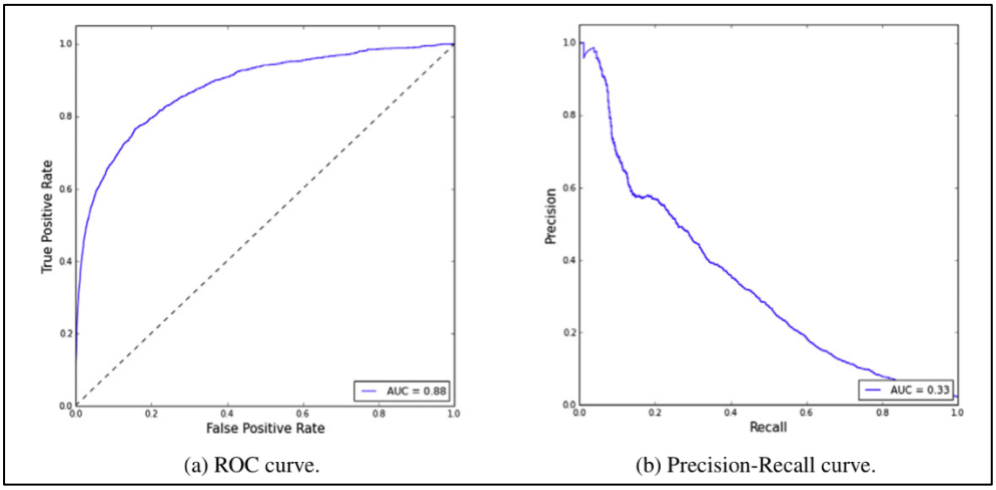

Fig. 4 Graphs showing the key performance indicators of the model (adapted from the case study [1])

The confusion matrix shown in Fig. 5 shows the performance of the model on the test set when the threshold level was chosen in accordance with the company's objectives of having only 20% of orders manually reviewed. The results show a high value of specificity at 98%, which is the fraction of legitimate orders which are approved. However, the recall value is rather low, as only 59% of fraudulent orders are actually refused. Overall, the combination of a random forest classifier with a manual revision of suspicious orders yields fairly good results that meet the targets set by the company.

	Legitimate (actual)	Fraud (actual)	Total
Legitimate (predict)	83441 (96.02%)	768 (0.88%)	83914
Fraud (predict)	1592 (1.83%)	1092 (1.25%)	2684
	Automation level:	80%	
	Recall:	0.587	
	Specificity:	0.981	
	Fallout:	0.019	
	Precision:	0.407	

Fig. 5 The confusion matrix of results when automatically approving 80% of records (adapted from the case study [1])

As society becomes more integrated with technology, more of the global economy will be transferred online. Therefore, online fraud is expected to continue to grow around the world. Fraudsters are utilising increasingly sophisticated methods and are targeting a growing number of industries and markets. Fraud detection techniques will also continue to develop. For example, 74% of UK executives are relatively convinced that biometrics will help to authenticate the vast majority of payments over the next 10 years [8]. Techniques such as this are crucial in gathering more information about a customer so that a better identification system can be created, this will make it harder for fraudsters to impersonate customers. However, data science techniques can only play a limited role in combatting cybercrime and online fraud. Governments have to invest in cybercrime units and work together in order to prosecute cybercriminals that operate internationally.

The internet is providing the perfect place for criminals to commit fraud. Data science techniques can be used to identify fraudulent transactions, as shown in the case study where a random forest model was implemented in the fraud detection system of a large online retailer. Data from the National Hunter Fraud Prevention Service showed that levels of fraud in the UK increased by 33% in May 2020, when criminals looked to profit from the disruption that the Covid-19 outbreak brought to both businesses and their customers. This highlights the relevance and importance of creating secure fraud detection systems. There is no permanent solution to fraud, however individuals, companies and governments can work together to reduce it.

Word Count: 1430

References

[1] Nuno Carneiroa, Goncalo Figueiraa and Miguel Costa, “A data mining based system for credit-card fraud detection in e-tail,” in Decision Support Systems, Vol. 5, 2017, pg. 91-101

[2] Button M., Lewis C. and Tapley J. “Not a victimless crime: The impact of fraud on individual victims and their families,” in 2014. [Online]. Available: https://doi.org/10.1057/sj.2012.11

[3] Adrian Banarescu, “Detecting and Preventing Fraud with Data Analytics,” in Procedia Economics and Finance, Vol. 32, 2015, pg. 1837-1836

[4] Anthony, “Why is it so Hard to catch cybercriminals?” 2020 [Online]. Available: https://blog.tmb.co.uk/why-is-it-so-hard-to-catch-cyber-criminals

[5] Ghosh and Reilly, "Credit card fraud detection with a neural-network," presented at the 1994 Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, HI, USA

[6] Tony Yiu, “Understanding Random Forest”, Towards Data Science, 2019, [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[7] Venkata Jagannath, “Random Forest Template for TIBCO Spotfire” 2020 [Online]. Available: https://community.tibco.com/wiki/random-forest-template-tibco-spotfire

[8] Author Unknown, “New Dimensions of Change: United Kingdom Report” TransUnion Information Group Limited, 2020 [Online]. Available: https://solutions.transunion.co.uk/digitalconsumertrust/