Analysis of Machine Learning techniques to detect credit-card fraud on e-commerce platforms

Assignment 3: Full Research Proposal

Harsha Lingutla (n9738355)

Problem Statement

Credit card fraud is increasingly affecting the revenue streams of many e-commerce retailers. "Recent estimates indicate that identity crime costs Australia upwards of $1.6 billion each year, with the majority (around $900m) lost by individuals through credit card fraud, identity theft and scams" (Australian Federal Police, 2019).

Primary credit card provider, MasterCard was one of the first companies to use Artificial Intelligence to help minimise credit card fraud (Felt, 2017). Through its Decision Intelligence platforms, MasterCard applied predictive analytics powered by machine learning to lower the number of false positives by 50% (Marr, 2018). Evidently, with a growing awareness of machine learning and artificial intelligence, retailers are now more willing to invest in technologies that mitigate credit-card fraud. However, as they often operate on budgetary constraints, financial return is a primary driver of future implementation. Due to the plethora of information available, it is vital to survey all existing methodologies.

Accordingly, this paper will compare the utility of some conventional credit card fraud detection techniques. Drawing on current scientific literature, these machine-learning techniques will be evaluated against cost, speed and accuracy considerations. Given multiple criteria are being consulted, organisations can select an appropriate technology based on their strategy (devise their weightings for each parameter). Ultimately, retailers with alarming levels of credit card fraud are provided with a convincing reason to invest in newer technology.

Research Question

Credit card fraud detection techniques are highly sought-after today. Of the several Machine Learning fraud detection techniques available, which is the best in terms of cost, speed, and accuracy considerations for large e-commerce retailers?

Definitions

- **Machine Learning:** A branch of Artificial Intelligence where systems can learn from data, identify patterns and make decisions with minimal human involvement.
- **Cost**: Refers to ease of implementation and use. They are sourced from employee qualitative feedback. Equally, a dollar amount (cost of asset + installation + maintenance) is determined.
- **Speed:** Processing speed (Low, Medium, High).
- **Accuracy:** Represents the fraction of the total number of transactions (both genuine and fraudulent) that have been detected correctly.
- **Error rate**: Shows how much the result value deviates from an actual value.
  Error Rate = 1 - Accuracy
- **Large organisation:** Any organisation that has a combined turnover greater than $250 million (ATO, 2018).

The research question directly **addresses the nature of the problem**. Recapping, the problem examines the most common fraud detection technologies available for e-commerce retailers. By conducting a performance analysis of these detection technologies, organisations can gauge their feasibility. Are they value-adding? A cost-benefit analysis can also address this question. Through data collection, an entity can identify what the strengths and limitations of each technology are. This highlights what the technology serves and its primary purpose.

Moreover, the question will enable retailers with problematic levels of credit card fraud, an authoritative reason to invest. Importantly, these results can be weighed against other literature surveys. Is the data analogous?

New knowledge is produced in the form of performance data. This quantitative data can be easily compared to different technologies. Additionally, qualitative data is drawn from the cost criteria. This ensures, a comprehensive, unbiased method of data collection. Given there are multiple criteria assessed, organisations can choose an appropriate technology based on their objectives (devise their weighting for each of the three criteria). Are they looking for technology which ensures total accuracy in fraud detection, or only interested in purchasing a low-cost solution? This gives the research question a far broader scope, as a broad audience is targeted.

Research Methodology

Overall Strategy

The process used to solve the research problem is not limited to one methodology.
Instead, a multi-faceted approach is employed.

Initially, **artefact design and testing** are used to develop and evaluate the Machine Learning techniques on the e-commerce site. The paper *Credit card fraud detection using machine learning techniques: A comparative analysis* follows a similar approach. Researchers investigated the performance of several techniques: naïve Bayes, k-nearest neighbour and logistic regression on highly skewed credit card fraud data. A dataset of credit card transactions is sourced from European cardholders containing 284,807 transactions (Awoyemi, Adetunmbi & Oluwadare, 2017). The work is simulated in Python, and the performance of the techniques are evaluated based on accuracy, sensitivity, specificity and precision criteria. Similarly, in this paper, cost, accuracy and speed dimensions will be evaluated.

To complement these findings, qualitative feedback is produced from user-completed **surveys**. These surveys will identify potential loopholes in the design of the site. Also, they can reveal if the prospective business is comfortable in operating the new algorithm.

Research Steps

Step	Description	Data Collection Approach
1	Review current machine learning algorithms and tools. Examine credible papers noting any discrepancies in performance and design.	Qualitative data is sourced from the *approach and conclusion* section of each paper.
2	After conducting a Literature Review, shortlist 3 of the best algorithms for fraud detection. This should be based on cost, speed and accuracy ratings.	Quantitative data will be collected from the *results* section of each paper. Performance metrics will include accuracy (error rate) and processing speed data.
3	Select an online transaction data set to implement the algorithms. There should be a minimum of 1000 unique records to ensure the validity of results. The data set should contain some (not restricted to) the following user information: credit card usage, purpose, the current balance in credit card, Holder of a credit card etc.	Integrate Kaggle transaction dataset. The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions (Kaggle, 2019).
4	The shortlisted (3) algorithms are implemented and tested in Python using their respective library and in-built packages. All algorithms need to follow the 70/30 rule. 70% of the entire dataset for training, and 30% for testing.	Automatic monitoring and logging of events within Python.
5	Implemented algorithms provided results on their accuracy, time duration (speed). The results should be visually represented (in clear graphs, tables), so the reader can quickly evaluate algorithmic behaviour.	Quantitative data is displayed on the analysis chart. The data is mined from Python, which logs performance data automatically.

6	Integrate top 3 Algorithms on e-commerce sites. Initially, a simulation should be developed to test various algorithms within an e-commerce system. *RetSim* is designed to be used in developing and testing fraud scenarios at a retail store while keeping business-sensitive and private personal information about customer's consumption secret from competitors and others (Lopez-Rojas, 2016). *RetSim* is the first simulator with the purpose of fraud detection on the retail store domain (Lopez-Rojas, 2016). This should be used as an excellent model to emulate. This will model the user's perspective of the website, once the algorithms are deployed. In this way, the prospective client is aware of the user's relationship with the site and can apply specific design strategies to prevent and detect credit card fraud.	Quantitative data (accuracy and speed dimensions). Data should then be extracted to a visualisation tool (Gephi). Results can be compared with similar peer-reviewed case studies.
7	Send follow-up surveys to potential users (e-commerce businesses who were part of step 6). How did different stakeholders (manager, employees) go with managing the new technology?	Developed in *Qualtrics Research Core*. Devise questions which generate feedback on ease of implementation and use topics (cost criteria).
8	Publish findings – the select best technique	Findings should be structured according to the requirements of a scientific paper. Quantitative data (in a visual format), should be supported by existing literature, and client's feedback.

Proposed Analysis Process

Quantitative analysis will form the basis of capturing the accuracy and speed dimensions of each algorithm. The information will be determined through a website simulation, and then extrapolated to a data visualisation tool, where it becomes understandable to the user. This new knowledge can be aligned to meet the unique objectives of each e-commerce platform. Is their solution centred around cost-saving or accuracy?

Similarly, surveys are a low-cost approach to see if the techniques are workable within e-commerce organisations. They can be employed after results from the simulation approach are known. For cost dimensions (ease of implementation and use) criteria, qualitative feedback can be sourced from surveys. Participants can be chosen from some of the significant e-commerce businesses. These surveys will compare the techniques and invite participants to suggest ways to improve the algorithms. Additionally, a dollar amount should be estimated from reputable sources.

Resource Estimate

- **Length of Project**: **9 - 12 months**. Maximum of 6 months assigned to develop and simulate algorithms on e-commerce sites accurately. Also, it may take considerable time to find willing clients.
- **No of Research Staff**: **6.** Split the research methods into three groups of 2. One group solely focuses on literature review and Python implementation. Another team specialises in website simulation. The final group creates surveys in Qualtrics software. Once the first stage is completed, team one can join the website simulation team. Also, all members should provide equal input to publish the research paper.
- **Special Resources/Equipment**

a) Access to Kaggle Transaction Data & Python Language

b) Gephi to design charts

c) Qualtrics Research Core for Survey generation

d) Examine RetSim (see page 5 for details)

e) Strong understanding of Machine Learning and Python programming required

Expected Tangible Outputs

Tangible Output	Link to Research Methodology	Deliverables
Data sets of shortlisted algorithms. Reveals the speed and accuracy of fraud detection.	Algorithms are implemented and tested in Python.	Events are automatically monitoring and logged within Python. This data is then extracted and visually presented in Gephi.
Data sets from an algorithmic simulation. Shows the practicality of the solution, in a commercial setting.	Algorithms are integrated onto e-commerce websites.	Data is extracted and presented in transparent charts using Gephi (Visualisation Tool).
Simulation software to test algorithms on an e-commerce platform	RetSim is used as an existing model to build on.	The software can be used for future research in the domain. For instance, the software can support simulations that examine different criteria.
Qualitative feedback on algorithms cost-effectiveness	Develop surveys in *Qualtrics Research Core.* The questions should focus on the user's ability to implement and manage the fraud-detection technology.	Findings are included in the published paper.

New Knowledge

Key questions answered by the end of the project.

- How effective are machine-learning algorithms at detecting fraud on e-commerce sites?
- What criteria is more valued by retailers? Based on survey results.
- What is the top-3 machine learning fraud detection techniques available?

New knowledge is produced in the form of performance data (accuracy and speed dimensions). This quantitative data can be easily compared to different techniques. Given multiple criteria are being consulted, organisations can choose an appropriate technique based on their own internal strategy (devise their own weighting of each parameter). Companies perplexed by the trade-off between accuracy and cost can now make a calculated decision based on the data presented. Each technique has notable advantages and disadvantages, which will be explored during the shortlisting phase. This information will enable retailers to pinpoint a technology which either is proven accurate or only a low-cost alternative. This gives this research question a far greater scope, as a broader audience is targeted. Also, through feedback raised in the surveys, one can pinpoint if there is any correlation between the performance data and organisation sentiment about the technique. In this way, the statistics on accuracy and cost can be given value and context. Ultimately, organisations are the ones who interact with technology. If it is not functional, implementation is unlikely.

After conducting a quantitative and qualitative analysis of the shortlisted techniques, the research problem is directly answered. Based on the accuracy, cost and speed criterion, the best machine learning technique to detect fraud is identified. Eventually, retailers with unhealthy levels of credit card fraud, are presented findings which encourage future investment.

References

J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," *2017 International Conference on Computing Networking and Informatics (ICCNI)*, Lagos, 2017, pp. 1-9. DOI: 10.1109/ICCNI.2017.8123782
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8123782&isnumber=8123766

ATO. (2018). Large business. Retrieved 10 September 2019, from https://www.ato.gov.au/business/large-business/

Australian Federal Police. (2019). Identity Crime. Retrieved 10 September 2019, from https://www.afp.gov.au/what-we-do/crime-types/fraud/identity-crime

Felt, C. (2017). Can AI Really Help Minimize Credit Card Fraud? - Clouded Blog. Retrieved 10 September 2019, from https://www.ibm.com/developerworks/community/blogs/ce53fcb5-53da-4efa-90e3-d92a77c52944/entry/Can_AI_Really_Help_Minimize_Credit_Card_Fraud

Kaggle. (2019). Credit Card Fraud Detection. Retrieved 13 October 2019, from https://www.kaggle.com/mlg-ulb/creditcardfraud

Lopez-Rojas, E. (2016). APPLYING SIMULATION TO THE PROBLEM OF DETECTING FINANCIAL FRAUD. *Blekinge Institution Of Technology Doctoral Dissertation Series*, *2016*(6).

Marr, B. (2018). The Amazing Ways How MasterCard Uses Artificial Intelligence To Stop Fraud And Reduce False Declines. Retrieved 10 September 2019, from https://www.forbes.com/sites/bernardmarr/2018/11/30/the-amazing-ways-how-mastercard-uses-artificial-intelligence-to-stop-fraud-and-reduce-false-declines/#290a458e2165

Reflective Statement

Honestly, the steps taken to complete this task was straightforward. As Assignment 3 followed a similar structure (content varied slightly), I completed the task in less time than before. Moreover, by attending lectures and diligently completing the preparatory tasks, I was not as flustered.

To complete the research steps, I consulted the lecture slide examples. This gave me a broad understanding. To better devise a research methodology, I also examined literature which concentrated on a similar research problem to mine. For the expected tangible outputs and new knowledge, I took my summary notes in Assignment 2 and expanded on them. Again, an additional literature survey allowed me to integrate concepts on newer technology (RetSim) to solve my research problem. Overall, I found the research process relatively linear; given the consistent effort, I have made since the beginning of the semester. However, my toughest challenge was to find areas of improvement, given that I received good marks in my previous two assessments. Mainly, I needed to ensure that my third piece was not a direct copy-paste.

In the future, especially in the lead up to master’s thesis units, I hope to maintain the good habit of starting my assignments as soon as they are released. This ensures quality content (more time to refine assignment) and more importantly prevents unwanted stress. I recognise the need to devote equal attention to all stages of the research process. They build on top of each other. Similarly, I should engage my tutor more frequently. This strengthens the notion that 'there is no stupid question', as they provide valuable guidance and tips (what am I doing well and what I need to work on). Equally, the writer's overconfidence is suppressed.