COS720 – Prof Jan Eloff

2017 – Assignment

Total marks: 50 with a bonus of 10 marks

Task completion:

1. **18 May 2017 - Submit on COS720 web portal and sign anti-plagiarism agreement.**
   a. **Upload all code/screenshots/visualisations of data/project design documentation**

2. **18 May 2017 - Book for a hands-on practical demonstration on 23 May 2017**

The history of Cyber Security goes back to approximately 1991 when the Computer Science and Telecommunications Board (CSTB) reported that security was the ''protection against unwanted disclosure, modification, or destruction of data in a system and also [to] the safeguarding of systems themselves." (CTSB, 1991) As the internet has progressed, so too has this definition to be more holistic. The following case study will test your logical application of this definition by following some data science steps. Please note that for this case study, you will focus your efforts on ensuring that the data science system that you create cannot be exploited by external attackers. This will be done, mainly by task 2.

Fraud costs the Insurance Industry R4bn a year. It would be greatly beneficial if we could predict insurance fraud using Big Data Science. In this assignment, you will be performing 4 main components of data science. A large part of data science is the fact that models created contain a substantial amount of information that cannot be leaked. The following tasks should get you thinking about data, its integrity and machine learning.

Follow the below steps. There are no right or wrong answers.

1. Data Generation and Exploratory Data Analysis (EDA):

From appendix A. Generate 100,000 fabricated insurance claims through any technique. If you would like, for ease of use, your names and surnames etc. can be random numbers or combinations of characters. Ensure that you have at least 100 fraudulent claims with reasons (these reasons can be the same, but try to have at least 5 different ones). These reasons can be generated from online research or through your own ideas. An example of a fraudulent insurance claim would be a person who has not paid any insurance premium but has claimed a substantial amount. **(10 Marks)**

Once the data has been generated perform Exploratory Data Analysis (EDA) on the full dataset. The idea here is to know what data you have available and the distributions of each of the attributes. Motivate whether you believe data scaling is required – this will be important during the machine learning step discussed later. **(10 Marks)**

2. PPDM:

It is illegal to use data in South Africa without the authority of the person to whom it is about (Luck, 2014); the data has to be anonymised first. Xu et al. (2016) maintains that it is possible to preserve privacy when performing machine learning by means of Privacy Preserving Data-Mining (PPDM). In-

vestigate techniques to anonymize the data you have generated in 1. Generate a script/algorithm to anonymize your data set. Please bear in mind that the data needs to maintain its machine learning value without having enough information to uniquely identify a person. Once you have done this, motivate your design choices. Ask yourself, if a system contains this data and it is attacked by an outside threat, will people's personal information be compromised? **(10 Marks)**

3. Data Cleaning:

To correctly predict Fraud, your data needs to be reliable. Dirty data is data which has values which are missing, incorrect or inconsistent (Krishnan et al., 2015). "Dirty" some records in your data-set and subsequently write an algorithm that cleans it. You can use existing algorithms, provided that you reference them where necessary. **(10 Marks)**

4. Machine Learning

Your last task is to run a supervised machine learning algorithm on your dataset to gain some knowledge. Your machine learning algorithm should have "Fraudulent claim indicator" as the dependant variable. Try and find a way using your machine learning algorithm to predict whether a new claim that comes in would be fraudulent. Motivate your machine learning algorithm choice. Also indicate which features deemed to be most valuable in your predictive model and why.

(10 Marks)

5. Bonus Question:

In this task, you need to write approximately a page on how you could apply the data science techniques that you have used in this insurance claim assignment, to cyber security in general (Steps 2 - 4). Although, the application of the aforementioned tasks are not specifically cyber-security focused, the techniques used can be very applicable and similar in predicting malicious attacks.

(10 Bonus Marks)

References

1. Luck, Russel, "POPI-Is South Africa keeping up with international trends?", May (2014).
2. Xu, Lei and Jiang, Chunxiao and Chen, Yan and Wang, Jian and Ren, Yong, "A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining", Computer (2016), 54--62.
3. Krishnan, Sanjay, et al. "SampleClean: Fast and Reliable Analytics on Dirty Data." IEEE Data Eng. Bull. 38.3 (2015): 59-75.
4. CSTB (Computer Science and Telecommunications Board). (1991) Computers at Risk: Safe Computing in the Information Age. Washington, DC: National Academy Press.

Appendix A

Example of a claim: For this research, you may generate only the highlighted fields (these are the most important)

Number	Field	Description
1	Fraudulent claim indicator	True or false indicator showing whether the claim was fraudulent or not
2	Fraudulent claim reason	Reason why claim was determined as fraudulent
3	Date Of Loss	Date that incident occurred resulting in loss for policy holder
4	Time Of Loss	Time that incident occurred resulting in loss for policy holder
5	Date Of Claim	Date that policy holder reported loss to broker/insurer
6	Agency/Broker Unique ID	Field to uniquely identify the broker
7	Insured Unique ID	Field to uniquely identify the Insurer
8	Insured's name	The name of the person who the policy is under
9	Insured's surname	The surname of the person who the policy is under
10	Policy holders telephone no	The telephone number of the policy holder
11	Age	Age of policy holder
12	Gender	Gender of policy holder
13	Kind of loss	Type of loss that occurred (Fire, theft etc.)
14	Address of incident	Address at which incident occurred
15	Address of policy holder	Residential address of policy holder
16	Police or Fire Dept To which Reported	Name/ Area of Police/ Fire Dept where injurious harm was reported
17	Policy holder Street	Street where policy holder lives
18	Policy holder Province	Province where policy holder lives
19	Policy holder City	City where policy holder lives
20	Policy holder Area	Area where policy holder lives
21	Policy holder Postal Code	Postal code where policy holder lives
22	Province	Province where loss occurred
23	City	City where loss occurred
24	Area	Area where loss occurred
25	Postal Code	Postal code where loss occurred
26	Marital Status	Marital status of policy holder
27	Date Of Birth	Date of Birth of policy holder
28	Sum Insured	The total value of what the policy holder is insured for
29	Probable Amount of Entire Loss	The estimated value by policy holder of the expense of the loss
30	Date of Settlement	The date that the claim was settled by insurer/ broker
31	Total Policies Revenue	The total premium received by the insurer from the policy holder

32	Amount Claimed	After evaluation, the value of the amount claimed
33	Amount Paid	The claim amount paid by the insurer to the policy holder
34	Policy Start Date	The start date of the policy holders insurance coverage
35	Policy End Date	The end date of the policy holders insurance coverage
36	Other party damage	The amount claimed from third party for incident
37	Other party name	Name of the other person if more than one person is involved in incident
38	Other party surname	Surname of the other person if more than one person is involved in incident
39	Other party insurer	The name of the insurer of the other party
40	Other party Street	Street where other party lives
41	Other party Province	Province where other party lives
42	Other party City	City where other party lives
43	Other party Area	Area where other party lives
44	Other party Postal Code	Postal code where other party lives
45	Assessor unique identifier	Unique identifier of the claims assessor
46	Total excess	The total excess paid by the policy holder
47	Type of insurance coverage	The type of insurance coverage eg comprehensive only
48	Title holders name	The name of the person who owns the property/ items damaged/lost
49	Title holders surname	The surname of the person who owns the property/ items damaged/lost
50	Short Description	Short description of the loss
51	Long Description	Long Description of the loss
52	Claim service provider	Service provider who will repair/ offset the damage