FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student:	Stanislav Němec
Oponent práce:	Mgr. Jan Starý, Ph.D.
Název práce:	Machine learning for financial crime detection
Obor:	Znalostní inženýrství

Datum vytvoření: 19. 6. 2019

Hodnotící kritérium:	*Způsob hodnocení – následující škálou 1 až 4:*
1. Splnění zadání	*1=zadání splněno,* *2=zadání splněno s menšími výhradami,* ***3=zadání splněno s většími výhradami,*** *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Komentář:

1. The survey of data mining techniques is done in chapter 5 (pages 15,16).
2. The implemented algorithms only work with accounts, not transactions.
3. The "improvements" meant to reduce false positives _raise_ false positives (from 90 to 370, 144, 116, respectively)
4. The "interpretability" of the models is never mentioned; we never see a single decision rule.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
2. Písemná část práce	*55 (E)*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3. Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Komentář:

Chapter 2 is an intro to the problem of money laundering, but makes a rather poor case for using machine learning:
"Current systems used by financial institutions are mostly based on rules assembled by consultants.
The rules are based on best practices and experience collected from the clients.
Having a fixed set of rules creates a risk of them being exposed and thus easily circumvented."
How exactly is machine learning better at that? In fact, 2.2.1 says
"Other important feature, that system for detecting suspicious transactions should have,
is an ability to provide a reasoning behind its decision. When producing a SAR,
there needs to be an explanation of why the activity is considered suspicious."
When using ML for these detections, the only explanation is "our model said so".

Chapter 3 is a concise intro to machine learning in general and evaluating the trained models.
Section 3.5 (Imbalanced classes) should probably mention that our intended classes
are extremely imbalanced, assuming the vast majority of transactions will be non-fraudulent.

Chapter 4 describes the dataset, obtained by a simulator:
a set of 500000 transactions among 10000 accounts over a period of 150 days.
The text later talks about "activities" and "entities" (meaning transactions and accounts, which is never stated).
The dataset contains zero-amount transactions (?), as well as e.g.

id,step,type,amount,nameOrig,oldbalanceOrg,newbalanceOrig,nameDest,oldbalanceDest,newbalanceDest,isFraud,alertID
0,1,CASH-IN,80.97,143,0.0,0.0,0,0.0,80.97,0,-1
1,1,CASH-OUT,59.29,0,0.0,0.0,185,0.0,59.29,0,-1

- how exactly do CASHIN and CASHOUT differ then?
(We never see any of these in the text.)

Chapter 5 surveys the existing solutions, and risk scoring is chosen as the method
Section 5.2 (Decision tree) describes the baseline model (the rest of page 17 being empty for some reason).
Note that the eventual decision tree is a finite set of yes/no questions; that's exactly
"a fixed set of rules which creates a risk of them being exposed and thus easily circumvented",
avoiding which was the whole premise.

Chapter 7 (Realization) describes first the preprocessing datasets.
According to 7.1, about 1500 records were transactions of zero amounts,
which makes the original simulation suspicious to say the least.
Why are we using a similator that produces zero transactions?

Only here do we learn that we will _not_ be using the given log of transactions:
they are aggregated as global stats by account number (total in, total out, num of transactions, average balance),
and we will only classify accounts, not transactions; unlike the original transaction log, this set of account aggregates
is _not_ available in plain csv form.

The set of accounts is split into a training set and a testing set, but we don't know how exactly.
The distribution of "target variable" in both parts (probably meaning the attribute of being fraudulent)
is stated to be "similar to the distribution in the original dataset", without saying what the distribution is.
Figure 7.1 provides a needless picture, as opposed to simply stating the numbers.

Table 7.1 (the confusion matrix) reports 265 true positives, 2370 true negatives, 90 false positives, and 257 false negatives.
After "improving" the decision tree, we get 370, 144, and 116 false positives, respectively.

We never see the eventual trained decision tree, i.e. the finite set of yes/no splitting questions,
or an example of a single one. The set of questions is _not_ available in plain form.
The overal table 7.6 only lists false positives, not false negatives.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
3. Nepísemná část, přílohy	*70 (C)*

Popis kritéria:
Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce
– opakovatelnost experimentů

Komentář:

The main result is a the trained decision tree, in the form of a Python Pickle.
The inputs are hardwired in the code.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Hodnocení výsledků, jejich využitelnost	*55 (E)*

Popis kritéria:
Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Komentář:

The model misses about 50% of the fraudulent accounts.
I don't think that would be acceptable in a bank.

Hodnotící kritérium:	*Způsob hodnocení – nehodnotí se*
5. Otázky k obhajobě	

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:

1. How many of the simulated accounts partake in fraudulent transactions?
2. Why do you only work with accounts as aggregates, not individual transactions?
3a. How large is the trained (baseline) decision tree, i.e. how many splitting questions?
3b. How exactly does the training decide what questions to use? Is there more than one for a given variable?
(How many questions do we have for e.g. the average balance, given that the question is simply an inequality?)
3c. Can you show a walk of an example account through the tree, for each of TP, TN, FP, FN?
3d. If it's not too large, can you show the rules in full?
4. What are the complete confusion tables for all the models? (Table 7.6 only lists FP)?

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Celkové hodnocení	*55 (E)*

Popis kritéria:
Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.

Text hodnocení:

Given the premise of the thesis, i.e. replacing expert-crafted decisions to reveal financial fraud
with machine learning, I would expect a thorough discusion of how exactly is the trained model
better suited for that, what the deciding questions are (!), with convincing examples.
Instead, we run a non-descript model over random data, with a 50% failure rate.

Changes to the tree, supposed to reduce false positives (one of the main goals),
in fact raise false positives - the text does not discuss that at all, or revise the
"improvements" in any way.

Podpis oponenta práce: