

Detecting fraud and outliers using R

Tomaž Kaštrun

BUDAPEST
October 25-27, 2016

BI FORUM

About



- BI Developer and data analyst (SQL Server, SAS, R, Python, C#, SAP, SPSS)
- 15years experience with MSSQL Server
- 15years experience data analysis and DM
- Working: Spar ICS Austria, Spar Slovenija
- MVP, MCT, MCPT, MCSE SQL Server
-   tomaz.kastrun@gmail.com
-  @tomaz_tsq
-  <https://tomaztsql.wordpress.com>
- Frequent community speaker at SQL and Microsoft events
- Blogger, Avid Coffee drinker, Bicycle junkie



My VISA account transaction overview

Moje kartice - arhiv prometa

Datum nakupa	Prodajno mesto	Znesek v originalni valuti	Znesek prometa v EUR
21.09.2016	WHO INTERNET		50,00
21.09.2016	WHO INTERNET		100,00
21.09.2016	WHO INTERNET		50,00
04.10.2016	WWW.ALIEXPRESS.COM	2,09 USD	1,88
			Skupaj v breme: 201,88
			Skupaj v dobro: 0,00
			Za plačilo: 201,88

Želim pregled prometa za kartico:

Visa 4106 6900 5046 4190 (Tomaž Kaštrun) ▼

Mesec: 18.10.2016 ▼

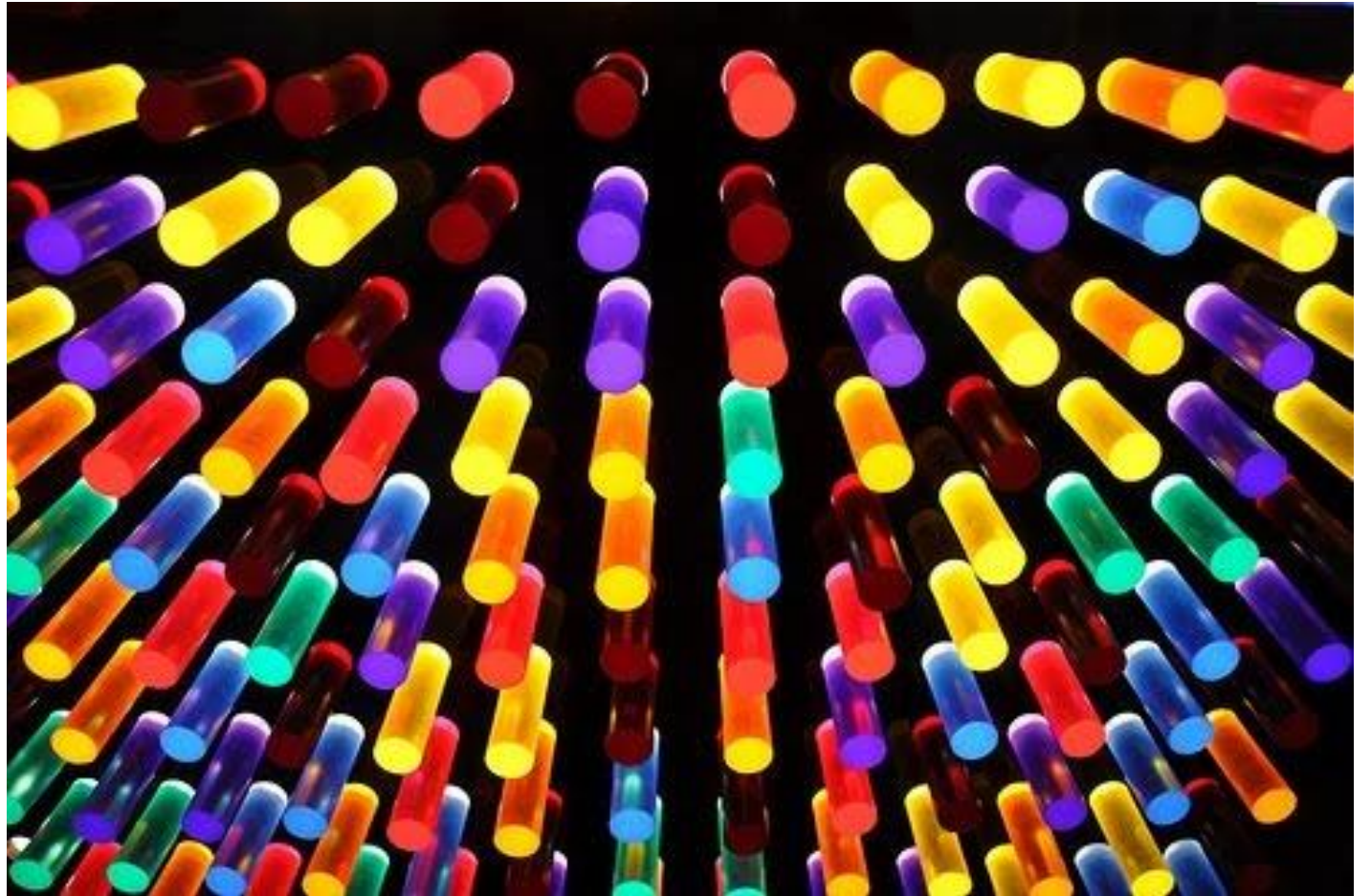
Noticed on 18.10.2016



Source: <https://klik.nlb.si>

 #budapestBI

Finding Patterns



Fitting right values



Applying mathematics



Pritožba imetnika plačilne oz. BA/Maestro kartice/zahtevek za fotokopijo potrdila o nakupu (slip)

Pojlite na naslov: Sektor spremljanja poslovanja s fizičnimi osebam
 Oddelček za kartične in bančnematne spremembe
 Faks: 01 476 54 31 na kopice naslov
 Faks: 01 476 54 32 za pravna oseba
 Smarčeva 130A
 1000 Ljubljana

Podatki imetnika plačilne kartice

Ime in priimek: DOMAZ HASTROW Davčna številka: _____
 Naslov: MAROLTOVA 4, 1000 LUBJANA Tel: 040 474 514
 Številka kartice: 4186 6900 5046 4190

Reklamacija nakupa/storitve

Prodatno mesto	WHO INTERNET	Datum nakupa	21.09.2016	Znesek nakupa	50.00 €	Kraj nakupa	INTERNET	Datum bremenitve	21.09.2016
Prodatno mesto	WHO INTERNET	Datum nakupa	21.09.2016	Znesek nakupa	50.00 €	Kraj nakupa	INTERNET	Datum bremenitve	21.09.2016
Prodatno mesto	WHO INTERNET	Datum nakupa	21.09.2016	Znesek nakupa	100.00 €	Kraj nakupa	INTERNET	Datum bremenitve	21.09.2016

Nakupi/storitve vključno s spletnimi transakcijami

- Zgoraj navedene transakcije nisem opravil jaz, niti nisem nikogar pooblašil, da opravi nakup namesto mene. Kartice nisem izgubil, niti mi ni bila ukradena, ves čas sem jo imel pri sebi. Kartico sem blokiral dne 18.10.2016
- Na zgoraj navedenem prodajnem mestu sem opravil nakup v višini _____ dne _____.
 spornega nakupa v višini _____ nisem opravil jaz. Kartice nisem izgubil, niti mi ni bila ukradena, ves čas sem jo imel pri sebi.
- Zgoraj navedeno transakcijo sem na prodajnem mestu opravil samo enkrat.
- Prodajno mesto je sprovedlo potrdilo o odobritvi (kreditni slip), vendar do danes za zgoraj omenjeni znesek moja kartica še ni bila odobrena. Prilagam dokazilo.
- Prodajno mesto je sprovedlo potrdilo o odobritvi (kreditni slip), vendar je bila moja kartica bremenjena namesto odobrena. Prilagam dokazilo.
- Moja kartica je bila obremenjena za znesek v višini _____ namesto _____.
 V prilogi vem pošiljam fotokopijo računa oziroma potrdila o nakupu.
- Sporna transakcija je bila plačana z drugim plačilnim sredstvom, in sicer _____, Prilagam dokazilo, da sem zadeve poskušal na prodajnem mestu urediti sam, ter dokazilo o plačilu.
- Zgoraj navedene transakcije ne priznam, saj na tem prodajnem mestu nisem opravil nakupa. Dne _____ sem kartico preklical zaradi (ustrezno obkroži): a) kraje, b) izgube, c) zlorabe, d) poneverbe.

Hotelske storitve

- Hotelsko rezervacijo sem pisno / telefonsko preklical dne _____ Preklicna številka je _____.
 Prilagam dokazilo, da sem zadeve poskušal na prodajnem mestu urediti sam, ter dokazilo o plačilu.
- Hotelsko rezervacijo sem pisno / telefonsko preklical dne _____, vendar preklicne številke nisem prejel. Prilagam dokazilo, da sem zadeve poskušal na prodajnem mestu urediti sam, ter dokazilo o plačilu.
- Pri rezervaciji hotela sem kot garancijo za rezervacijo navedel številko kartice. O tem, da mi hotel zaračuna stroške nočitve za en dan, če ne odpovem rezervacije, nisem bil obveščen. Prilagam dokazilo, da sem zadeve poskušal na prodajnem mestu urediti sam, ter dokazilo o plačilu.
- Nikoli nisem uporabil storitev v navedenem hotelu, prav tako nisem opravil nobene hotelske rezervacije.

Obrnite!

Qm: PUK-20 (junij 2012) Novi ljubljanske banke d.d., Ljubljana, Trgovske 2, 1000 Ljubljana

Bankomat!

- Bankomat mi želenega zneska, za katerega je bila bremenjena moja kartica, ni izplačal.

- Izplačan je bil le delni znesek dviga na bankomatu v višini _____ namesto _____
 za kolikor je bila moja kartica tudi bremenjena.

Zahtevek za kopijo potrdila o nakupu (slipa) za zgorej omenjeno transakcijo (obkroži!):

- a) z bremenjenim zneskom se ne strinjam
- b) ne prepoznam transakcije
- c) za lastno evidenco
- d) zaradi suma zlorabe
- e) drugo

Drugi razlogi za zavrnitev s kratkim opisom

KREDITNE KARTICE EDNO NE UPORABLJAM, RAZEN
 V PRIMERU POTOVANJ ALI NAKUPOV PREKO AMAZONA,
 ALI EXPRESSA. DNE 21.09.2016, KO SO BSE ZSDILE
 ZLOABNE (FRAUD) TRANSAKCIJE, SEM BIL V SLOVENIJI/LJUBJANI/
 NA POSEBEN NAPOVEDAN UPORABLJAM VSAK DAN MAESTRO
 ZA PLAČILA. NA TA DAN, 21.09.2016, SEM OPRAVIL NAKUPE
 Z MAESTROM (SPAR, BENCINSKI SERVIS), A NOBENIH SPLETNIH
 NAKUPOV, Z VISA KARTICO.
 NAKUP (OZ. BREMENITVE) NA VISA KARTICI NISEM OPRAVIL
 JAZ IN SEM OPAZIL ŠELE DANES - 18.10.2016 - KO SOMI TRGALI
 17 TRZAJA ZNESEK V VIŠINI 200€.

Priloge

Imetnik v primeru ne upravičene reklamacije soglaša z bremenitvijo, v skladu s trenutno veljavno tarifo banke.

 Kraj in datum: LJUBJANA, 18.10.2016

 Podpis imetnika: 
Izpolni banka

 Enota banke: 019 NOVE LARSE

 Telefon: 01/587-48-88

 Faks: 01/5217-214

elektronski naslov: _____

 Kraj in datum: Lj, 18.10.2016

 Žig in podpis
 pooblaščenca banke izdajateljice /


#budapestBI

Moje kartice - arhiv prometa

Datum nakupa	Prodajno mesto	Znesek v originalni valuti	Znesek prometa v EUR
21.09.2016	WHO INTERNET		50,00
21.09.2016	WHO INTERNET		100,00
21.09.2016	WHO INTERNET		50,00
04.10.2016	WWW.ALIEXPRESS.COM	2,09 USD	1,88
			Skupaj v breme: 201,88
			Skupaj v dobro: 0,00
			Za plačilo: 201,88

Želim pregled prometa za kartico:

Visa 4106 6900 5046 4190 (Tomaž Kaštrun) ▼

c: 18.10.2016 ▼

Patterns

Right values

Mathematics



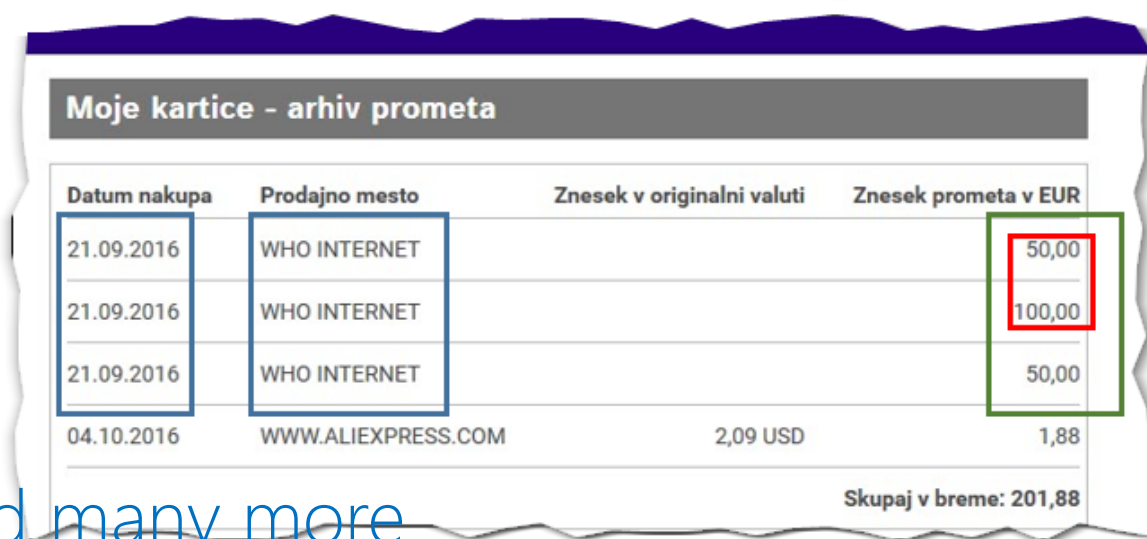
Finding what is „strange“

- „Strange“ values can:
 - Simply be out of context
 - Do not comply with business rules
 - Reject mathematical and/or statistical rules
- Methods for „finding“ values:
 - Semi-graphical methods
 - Uni-/Multi- variate statistical methods
 - Machine learning methods



Building rules


- Statistical rules:
 - Skewness, Kurtosis, Mean, SD and many more
 - Coefficients, Residuals, Loadings
- Procedural rules:
 - Rejecting all „strange“
 - Con-sequential rejection
- Learning rules:
 - AUC, precision
 - Supervized learning methods
 - Refining the models



Datum nakupa	Prodajno mesto	Znesek v originalni valuti	Znesek prometa v EUR
21.09.2016	WHO INTERNET		50,00
21.09.2016	WHO INTERNET		100,00
21.09.2016	WHO INTERNET		50,00
04.10.2016	WWW.ALIEXPRESS.COM	2,09 USD	1,88
			Skupaj v breme: 201,88

Rules example

- Statistical rules
 - Flag if out of $\pm 3SD$
 - Flag if spike is at the end of tail
 - Flag if more than 3 transactions ($> 50\text{€}$) occur within 5 minutes
- Procedural rules
 - Flag all that come out of „WHO INTERNET“ point of Sales
 - Flag if $x \bmod 10 = 0$
 - Flag if Ratio of Max and 2nd highest $> x$
- Learning rules
 - Flag if $(x \mid y \ \& \ z) > 0.4256$ AND $(y \ \& \ x) = \text{„WHO INTERNET“}$



The screenshot shows a transaction history table with the following data:

Datum nakupa	Prodajno mesto	Znesek v originalni valuti	Znesek prometa v EUR
21.09.2016	WHO INTERNET		50,00
21.09.2016	WHO INTERNET		100,00
21.09.2016	WHO INTERNET		50,00
04.10.2016	WWW.ALIEXPRESS.COM	2,09 USD	1,88

Skupaj v breme: 201,88

Annotations: A blue box highlights the first two columns (Date and Location) for the first three rows. A red box highlights the 'Znesek prometa v EUR' column for the first two rows. A green box highlights the 'Znesek prometa v EUR' column for the first three rows.

Stopping „strange“ values

- Outliers are:
 - Business-wise defined; with-in business there are many
 - Have no definite answer
 - Statistically/Mathematically it is not always that straightforward
- Start arbitrary
 - Check the data
 - Mark the outlier|strange|fraud
 - Build a model based on rules and knowledge-base
 - Learn from it



Data prep's

- Data preparing takes 50% of the time
- Problem with data is low presence of „strange“ data
- Sampling data (over or under sampling)
- SCD and RDBM issues (many to many, one to many relations, SCD I, II, III)
- Common sense; Day of week, holidays, Time of the day, Demographics, etc.



DEMO | BI FORUM



Machine learning approaches

- Detecting „strange“ values with:
 - GLM
 - Naive bayes
 - Random forest
 - Neural Networks
- Other techniques:
 - Density estimation
 - Time series
 - others



Measuring Efficiency and robustness

- Measure model efficiency:
 - Lift Chart
 - AUC
 - Accuracy, Precision, Recall
- Fraud detection is always work in progress and never 100% automated – remember: rules change, business change and people „frauding“ you also learn (change)



DEMO | BI FORUM



Key Take-aways

- Is a work in progress and constant learning cycle
- Combine statistical and mathematical methods for detecting and preventing „strange“ observations
- Always refine the models and look for false positives
- Rules are business-wise different between each others
- No definite definition of what Fraud is and no dedicated program.



Q&A

Demo data and R Code:

https://github.com/tomaztk/Detecting_outliers_and_fraud_BBI2016



<http://tomaztsql.wordpress.com>



tomaz.kastrun@gmail.com



@tomaz_tsql



/in/tomaztsql



<https://github.com/tomaztk>



BUDAPEST
October 25-27, 2016

BI FORUM



#budapestBI