Calibração de Threshold

Dezembro, 2017

Know the unknown..

oi

weDO technologies

AGENDA

1. Contexto
2. Necessidades
3. Abordagem
4. Âmbito / Fora do âmbito
5. Entregáveis
6. Plano
7. Proposta Económica

Calibração de Threshold

Contexto

De acordo com a acta de reunião produzida pela equipa Analytics, segue a respectiva proposta no contexto da Novo Antifraude – RAID FMS Onda 2 – TV Offline para a calibração dos pesos de acordo com as dimensões, periodos e contadores referidos no SDD - 2.3.3 RQN09 – Criar motor de score de risco de fraude da versão 1.02 do SDD SDD_BRA16POI03900_001_B-DSOL-PRJ25351 - RAID FMS - Onda 2_20171108_WeDo.

Calibração de Threshold

Necessidades

Para dar início aos respectivos desenvolvimentos analíticos, deverão ser garantidas as seguintes condições:

- Historico dos contadores
- Histórico de propostas
- Ambiente de desenvolvimento analítico
- Ligações aos ambientes

Calibração de Threshold

Necessidades – Histórico dos contadores

- Os modelos analíticos irão ser desenvolvidos de acordo com a informação histórica dos contadores de forma a poderem “aprender” com o passado e, de acordo com essa aprendizagem, irão devolver os pesos calibrados em conformidade.
- Desta forma, é necessario assegurar à equipa Analytics, a disponibilização da tabela **FMS_T_PREV_CONTADORES** num formato similar à tabela **FMS_T_PREV_MODEL_RESULT**, considerando as mesmas dimensões, períodos e contadores que as descritas no SDD - 2.3.3 RQN09 – Criar motor de score de risco de fraude da versão 1.02 do SDD SDD_BRA16POI03900_001_B-DSOL-PRJ25351 - RAID FMS - Onda 2_20171108_WeDo.
- Essa nova ta… …ayout:

Coluna	Descritivo
MESANO_ID	Mês e Ano a que se referem os dados
TIPO_CHAVE	Os mesmos que os usados na tabela FMS_T_PREV_MODEL_RESULT (CPF_CNPJ, CEP, etc.)
PERIODO	Os mesmos que os usados na tabela FMS_T_PREV_MODEL_RESULT (0a3M, 4a6M, Total)
FRAUDE_TOTAL	Fraude total observada no mês MESANO_ID inclusive
…	Os mesmos contadores que os definidos na FMS_T_PREV_MODEL_RESULT

Calibração de Threshold

Necessidades - Histórico de Fraude

- O histórico do evento alvo de análise deverá ser de igual forma assegurada para um período mínimo de 6 meses.
- Esta informação será derivada do atual contador **FRAUDE_TOTAL**.
- O contador **FRAUDE_TOTAL** será a variável target dos modelos analíticos a serem desenvolvidos. Isto é, assume-se que o contador **FRAUDE_TOTAL** traduz de forma fiável o resultado da análise final de uma proposta, isto é negada ou aprovada.
- Assim, o contador **FRAUDE_TOTAL** irá assumir os seguintes valores:
  - F R A U D E _ T O T A L =1 ⇔ Cliente cometeu algum tipo de fraude
  - F R A U D E _ T O T A L =0 ⇔ Cliente não cometeu fraude
  - F R A U D E _ T O T A L = *Null* ⇔ Não existe informação

Calibração de Threshold

Necessidades – Ambiente de desenvolvimento analítico

Dado o volume de dados a ser tratado no âmbito dos desenvolvimentos analíticos, é necessário assegurar à um ambiente analítico no qual seja possivel aplicar os diferentes algoritmos de machine learning.

A equipa Analytics sugere:

- Instalação do RSTUDIO no server do Client
- Instalação do Spark Yarn com Cluster StandAlone com 8 cores e 90GB

SW	Version	License Type
R	3.4.1	https://www.r-project.org/Licenses/
RStudio	1.1.383	https://www.gnu.org/licenses/agpl-3.0-standalone.html
Sparklyr packages	Dplyr_0.7.4 Sparklyr_0.6.4	https://www.gnu.org/licenses/agpl-3.0-standalone.html
Spark	2.2.0	https://www.apache.org/licenses/LICENSE-2.0
Hadoop (when used)	Cloudera 5.12	N/A

- Este ambiente pode ser criado em desenvolvimento ou em produção. Se for criado na maquina do Cliente poderá ser re-utlizado pelo mesmo considerando um mãximo de 3 utilizadores.

Calibração de Threshold

Necessidades – Ligação aos ambientes

Os desenvolvimentos serão realizados de forma remota, pelo que deve ser assegurado:

- Ligação remota à BD onde residem os historicos de informação solicitados nos pontos anteriores
- Ligação remota ao ambiente de desenvolvimento analitico referido na presente proposta

Calibração de Threshold

Abordagem

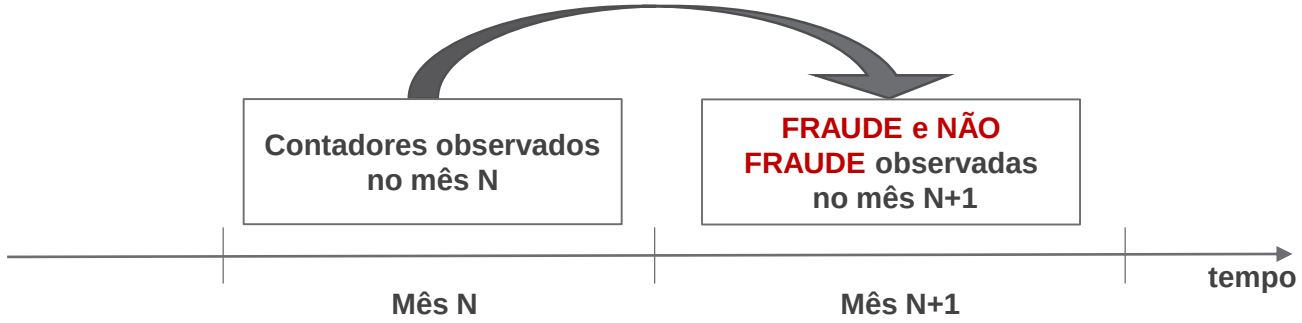
- Os modelos serão desenvolvidos de acordo com a variável target **FRAUDE_TOTAL**.
- Sendo que a variãvel target **FRAUDE_TOTAL** será a soma deriva do processo de derivação e de detecção
- Serão aplicadas diferentes tecnicas de DataMining que permitirão à obtenção dos pesos de modo a minimizar os seguintes erros:
  - **Erro tipo 1** – A proposta foi identificada como sendo **Fraude** apesar da indicação em contrario dada pelo modelo, isto é que não se iria converter numa Fraude
  - **Erro tipo 2** – A proposta foi identificada como não sendo uma Fraude apesar da indicação em contrario dada pelo modelo, isto é que se iria converter numa **Fraude**

Predicto (Modelo)	***Observado (Real)***	
	FRAUDE_TOTAL=0	**FRAUDE_TOTAL=1**
FRAUD_TOTAL=0	N.A	**Erro tipo 1**
FRAUD_TOTAL=1	**Erro tipo 2**	N.A

Calibração de Threshold

Abordagem

- Os modelos irão aprender com base na informação histórica dos contadores, para definir os que mais se relacionam com o comportamento de **FRAUDE** / **NÃO FRAUDE .**



Calibração de Threshold

Âmbito / Fora do Âmbito

Âmbito

- Modelos supervisionados para a calibração dos pesos descritos em no SDD - 2.3.3 RQN09 – Criar motor de score de risco de fraude da versão 1.02 do SDD SDD_BRA16POI03900_001_B-DSOL-PRJ25351 - RAID FMS - Onda 2_20171108_WeDo.
- As calibrações terão por objectivo indicar a importancia de cada um dos contadores para a tomada de decisão de aprovar ou não uma nova proposta de venda do serviço Oi TV.

Fora do Âmbito

- Geração da nova tabela referida Histórico de contadores
- Desenvolvimento de mecanismos de Calibrações automática dos pesos
- Qualquer intervenção no RAID*

*Qualquer intervenção que seja necessária no RAID será acautelada pela equipa de integração

Calibração de Threshold

Entregáveis

- Envio das calibrações para actualização da tabela dos pesos em RAID
- Código fonte do desenvolvimento dos modelos e obtenção dos respectivos pesos
- Relatório com o descritivo dos resultados
- Documento funcional inerente aos procedimentos internos da Oi

Plano

	Semanas						
	1	**2**	**3**	**4**	**5**	**6**	**7**
SetUp do Ambiente, preparação dos dados e amostragem							
Modelos analíticos (18 modelos no total)							
Support to Model deployment (Envio dos pesos calibrados)							
Report							
Presentation							

- Projecto com duração de 35 dias.

THANK YOU

Know the unknown...

weDo technologies