Machine Learning Engineer Nanodegree

Fraud detection in the Popular Pharmacy Program

Arthur Jahn Sturzbecher August 17, 2019

Proposal

This proposal presents a very domain specific problem related to the identification of frauds in the Popular Pharmacy Program of the Brazilian Ministry of Health. The next topics will dig into the domain, problem and possible solutions.

Domain Background

The logistics related to the distribution of supplies is a chalanging problem faced by any organization, when it comes to medicine, this chalange also brings the cost of people's lives. Designing a program that could distribute medicine to an entire nation as large as Brazil is a complex task and the Popular Pharmacy Program tryies to reach every corner of the country by refunding registered pharmacies and drugstores their costs with a list of medicines subsidized by the governmentto provide faster distribution by delegating such logistic to the drugstores' network. This facility comes with a cost: the number of frauds has increased in the last years due to the small number of public servers supervising the process of refund and the large number of registered pharmacies.

The process happens in the following steps: The pharmacy sells a medicine to someone; the seller provides important data to the drugstore, such as age, desease, prescription, and the ID number; The registered pharmacy submits the list of documents and follows some requirements to request a refund for the medicine; the Health Ministry verifies the provided data and if everything is ok, refunds the value provided in the invoice submitted with the request.

The frauds in the Popular Pharmacy Program happen in some ways, such as: providing fake pacient information; Registering multiple medicines to a pacient that has only requested to one medicine; Selling medicine to dead people and similar ones. They are all related to user's information that are in someway corrupted or missused to deceive the control program. There is a large set of information that could help in the identification of fraudulent transactions, such as the client location, the average age range of affected people, and many other indirect information that we can collect to provide as features to our algorithm.

Problem Statement

To simplify the problem to a reasonable and executable task, we have a set of transactions requested by each drugstore and we want to find the ones with higher chances of being fraudulent transactions. This is similar to a credit card fraud verification, that we can use as inspiration. The problem here also comes with a degree of business inteligence need, so the solution provided might be comprehensive.

Datasets and Inputs

There are two datasets provided by this task. The first one is a large amount of transactions that are not classified in fraudulent or not. The seccond is a small set of data that was manually analysed and has the label corresponding to each transaction.

The smaller dataset has 4000 registers of transactions that were classified as fraud, this dataset has only two attributes:

Transaction specific attributes:

```
 CO_SEQ_DISPENSACAO: Sequencial code for the requested transaction related to the pharr
BL_FRAUD_TRANSACTION: Boolean indication if the transaction was fraudulent or not
```

The larger dataset provided has aprox. 200.000.000 registers of transactions executed from 2013 to 2019 with 72 attributes of more than 2000 pharmacies registered in the program. The data has been anonimized in orther to remove sensible information. The provided dataset contains the following list of attributes that we can analyse:

Transaction specific attributes:

```
 DT_DISPENSACAO: Date of the requested transaction in the format "DD/MM/YYYY - HH:MM:S
CO_SEQ_DISPENSACAO: Sequencial code for the requested transaction related to the pharm
QT_DISPENSACAO: Quantity of medicine sold by the transaction
VL_UNITARIO: Unitary value of the medicine sold
VL_REFERENCIA_POPFARMA: Unitary value that goverment pays back for that medicine
```

Stablishment attributes:

```
 CO_SEQ_ESTABELECIMENTO: Pharmacy stablishment code
```

ICD attributes (International Statistical Classification of Diseases and Related Health Problems):

```
 NO_CID: Number of the ICD register
CO_CID: Code of the related desease
ST_REGISTRO_ATIVO_CID: Identifier if the ICD register is active
```

Pacient attributes:

```
 ST_VIVO: Boolean if the pacient is alive or not
CO_PESSOA: Identification of the pacient
DT_NASCIMENTO: Day of Birth
DS_MES_NASCIMENTO: Month of birth
CO_ANO_NASCIMENTO: Year of birth
SG_SEXO_PACIENTE: Gender of patient on birth
```

Patient's geoinformation attributes:

```
CO_MUNICIPIO_IBGE_PAC: Code of the city
NO_MUNICIPIO_PAC: Number of the city in Annual Commerce Research (ACR)
CO_AGLOMERADO_URBANO_PAC: Code of the urban conglomerate in ACR
NO_AGLOMERADO_URBANO_PAC: Number of urban conglomerates in ACR
CO_MACRORREGIONAL_SAUDE_PAC: Code of health macro region of health
NO_MACRORREGIONAL_SAUDE_PAC: Number of the health region
CO_MESORREGIAO_PAC: Code of the ACR meso region (meddium size region)
NO_MESORREGIAO_PAC: Number of the ACR meso region (meddium size region)
CO_MICRORREGIAO_PAC: Code of the ACR micro region (small size region)
NO_MICRORREGIAO_PAC: Number of the ACR micro region (small size region)
CO_MICRORREGIONAL_SAUDE_PAC: Code of the ACR micro health region
NO_MICRORREGIONAL_SAUDE_PAC: Number of the ACR micro health region
SG_UF_PAC: State initials
NO_UF_PAC: State number
NU_ALTITUDE_MUN_PAC: Geolocation altitude
NU_UF_LONGITUDE_PAC: Geolocation longitude of the state
NU_LONGITUDE_MUN_PAC: Geolocation longitude of the city
NO_REGIAO_SAUDE_PAC: Number of the health region in ACR
```

Pacient's goverment program participation:

```
ST_PARTICIPA_POPFARMA_PAC: Indication if the pacient benefits from Popular Pharmacy pr
```

Medicine attributes:

```
NO_PRODUTO: Number of the product
CO_PRODUTO: Code of the product
NU_CATMAT:
CO_PATOLOGIA: Pathology related to medicine
QT_MAXIMA: Maximum quantity of the medicine
QT_USUAL: Usual quantity of the medicine
CO_PRINCIPIO_ATIVO_MEDICAMENTO: Code of the active principle of the medicine
```

Producer attributes:

```
NO_FABRICANTE: Number of identification of the producer
NU_REGISTRO_ANVISA: Number of the producer National Sanitary Surveillance Agency regis
```

Stock control and price attributes:

```
CO_GRUPO_FINANCIAMENTO:
DS_GRUPO_FINANCIAMENTO:
VL_PRECO_SUBSIDIADO: Price specified by the government
QT_PRESCRITA: Prescription ammount
QT_SOLICITADA: Requested ammount
QT_ESTORNADA: Reversed ammount
```

Popular Pharmacy Program control attributes:

```
NU_LINHA_CUIDADO:
DS_PROGRAMA_SAUDE:
SG_PROGRAMA_SAUDE:
TP_PROGRAMA_SAUDE:
ST_PARTICIPA_POPFARMA_EST:
ST_PART_FARMACIA_POPULAR_EST:
QT_POPULACA_PORTARIA_1555_2013:
```

Additional Pharmacy geolocation attributes: (Same as pacients' attributes)

```
CO_AGLOMERADO_URBANO_EST:
NO_AGLOMERADO_URBANO_EST:
CO_MACRORREGIONAL_SAUDE_EST:
NO_MACRORREGIONAL_SAUDE_EST:
CO_MESORREGIAO_EST:
NO_MESORREGIAO_EST:
CO_MICRORREGIAO_EST:
NO_MICRORREGIAO_EST:
CO_MICRORREGIONAL_SAUDE_EST:
NO_MICRORREGIONAL_SAUDE_EST:
SG_UF_EST:
NO_UF_EST:
NU_ALTITUDE_MUN_EST:
NU_UF_LONGITUDE_EST:
NU_LONGITUDE_MUN_EST:
NO_REGIAO_SAUDE_EST:
```

The dataset was provided in partnership between theMinistry of Health and the Medicine Faculty Foundation in terms of the agreement 857860 published in the Official Diary of the Union. The dataset is not yet publicly available, but is accessible in a hosted GreenPlum database.

Solution Statement

The solution proposed is to apply a set of four unsupervised learning algorithms to identify possible frauds. According to this study, the set of algorithms to use for unsupervised fraud detection can be: DBSCAN; MeanShift; Gaussian Mixture Model (GMM); One-class SVM; Z-Score and Median absolute deviation (MAD); Hierarchical clustering; Hidden Markov Model (HMM); and Self-Organizing Maps (SOM). Since my intention is to focus on implementations provided by SciKit-learn, the selected models are: DBSCAN; Gaussian Mixture Model (GMM); Hierarchical clustering; and One-class SVM.

Benchmark Model

The benchmark model will be the study realized by Rémi Rodrigues that describes measurements and results related to using unsupervised learning for fraud detection. Also, the evaluation pretended here is simillar to the one made by Rodrigues, since we have a small dataset of labeled transacions that will be used to evaluate how well each algorithm perform. For DBSCAN the benchmark will use Euclidean distance, for Hierarchical clustering, GMM and One-class SVM the benchmark will be based on Gaussian kernel.

Evaluation Metrics

The metrics used will be Silhouette Coefficient and Quantization error for Unsupervised learning analysis and precision, recall and F1 score using the small dataset that is labeled.

Project Design

- **Programming language:** Python 3.6
- **Library:** Pandas, Numpy, Scikit-learn
- **Workflow:**
    - Establish basic statistics and understanding of the dataset; perform basic cleaning and processing if needed.
    - Train the Clustering models on the given data as-is to gauge the performance.
    - Fine tune the model's hyperparameters.
    - Perform training.
    - Perform individual benchmarks for each model.
    - Perform Comparative analysis between models.

References

1 - DOMINGUES, RÉMI. Machine Learning for Unsupervised Fraud Detection, 2015. Available in http://www.diva-portal.org/smash/get/diva2:897808/FULLTEXT01.pdf.

2 - PIERRE, RAFAEL. Detecting Financial Fraud Using Machine Learning, 2018. Available in https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9

3 - FREI, LUKAS. Detecting Credit Card Fraud Using Machine Learning, 2019. Available in https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8