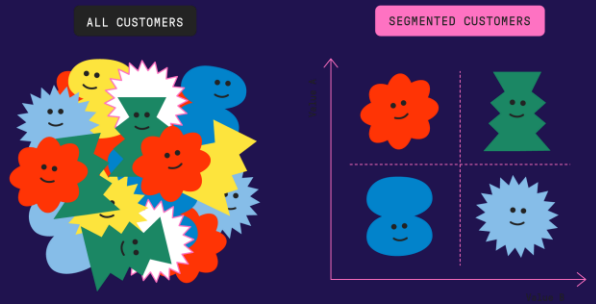
Application Banking Domain

Benazir de la Rosa
Data Science Consultant



Fraud Detection

Customer Segmentation

The demo will be divided into two parts:

- Part 1: Fraud Detection
- Part2: Customer Segmentation

Part 1: Fraud Detection

Agenda

01 | Business Case

02 | Dataset Description

03 | Descriptive Analytics

04 | Feature Engineering

05 | Deep Learning Model and Results

01

Business Case

Banking Domain

Business Case

The procedure to generate bank fraud applications goes more far away than just a machine learning/deep learning model. It includes hard rules, network checks, software checks, until it reaches the data science model.

The scenario presented in this use case simulates what you can do with bank domain data to detect fraud.

Fraud can happen anywhere and the most important key to create a good fraud application is to know the business and technology with its pros and cons. In such way that you can find clouds in the procedure and detect fraudsters.

02

Dataset Description

Banking Domain

Dataset Description

The data presented in this small demo was found online. The attributes are anonymize. Thus, the attribute definition is simulated.

There were two datasets used in this demo:

Info_01	Info_02
17286 Registers	17286 Registers
v6: type float	id: type integer
id: type integer	id2: type integer
id2: type integer	v1: type float
v4: type float	c1: type integer
v5: type float	v2: type float
	v3: type timestamp

Data cleaning findings:

- The timestamp had ~10% missing data
- The pair columns “id” and “id2” had same information
- Duplicate registers
- The dataset had no imbalance data

03

Descriptive Analytics

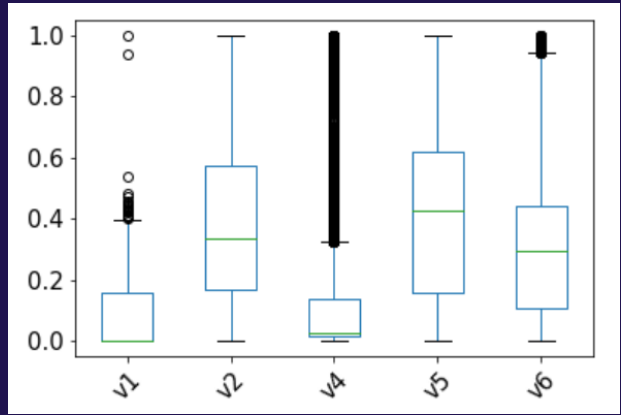
Banking Domain

Descriptive Analytics

The dataset had 6 days of data, ranging from the 2015/02/04 – 2015/02/10. (v3)

In general the attributes provided the next information:

- v1: It seems to be the amount of money involved in a transaction.
- v2: it is an indicator 19.0 – 23.0.
- v4: Total value in account. The population had a lot of people outstanding in terms of savings.
- v5: Ages of customers
- v6: Bank indicator.



Overall, we can say that the outliers available in v1, v2 and timestamp might be good to identify fraud.

04

Feature Engineering

Banking Domain

Feature Engineering

The following decision were taken with respect to this part:

- Time domain: the date was decompose and only day, hour, minute and second was use. The lack of data lead to discard month.
- Account transactions: To try to identify any unusual pattern in movements, I decided to create a new attribute to check the difference between the total balance and the transaction amount.

It was not possible to create more features due to the lack of information.In a real world application, you can track more patterns to identify anomalous behavior. Among them are the average transaction over a time window period, logins from apps/websites, tracking of atm visits, amount of times that a customer checks the balance, amount of transaction per time window period, geolocation from customer transaction etc.

05

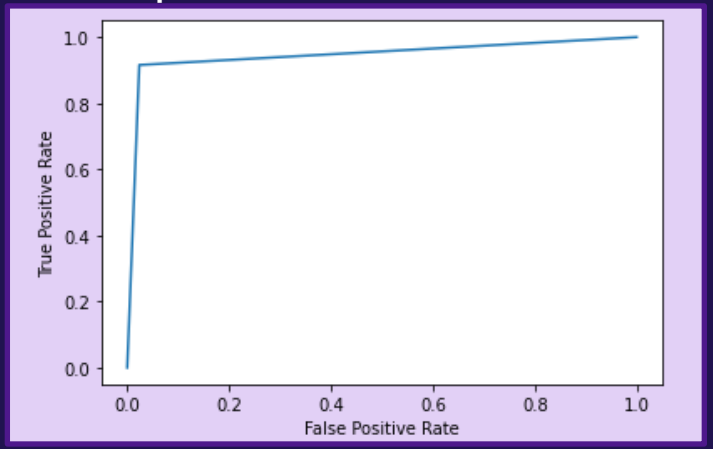
Deep Learning Model and Results

Banking Domain

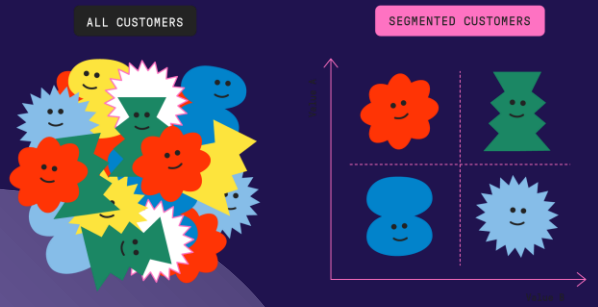
Deep Learning Model and Results

The selected model to predict the fraud was a feed forward neural network. A parameter tunning process took place with grid search. The parameters selected were the next ones:

- The neural network was created with two layers
- No dropout layer was use
- Optimizer: adam
- Batch size: 10
- Number of epochs: 20



The model was tested with a 10-fold cross validation and the metric chosen to predict its performance was accuracy. The model gave in average 94% accuracy.



Part 2: Customer Segmentation

Agenda

01 Business Case

02 Feature Engineering

03 Machine Learning Model and Results

04 Customer Descriptions

01

Business Case

Banking Domain

Business Case

The scenario presented in here ilustrates how a company can use users data to understand them better.

Customer segmentation is one of the first strategies use to leverage product selling.

Nowdays it is a key requierement to understand customers needs. The customer segmentation can be use to improve the next situations at a company:

- Detect good and bad payers
- Detect types of transportation paths
- Be able to create novelty offers which turn out into a successful revenue
- Detect at what hour might be better to call them
- Detect what kind of approach should you use with them according to their profile to sell them a product
- Detect marketing strategies by geolocation area

The list can continue. It is up to the company to choose their path.

02

Feature Engineering

Banking Domain

Feature Engineering

For this type of scenario, the features use are the same as the previous application.

- Decomposition of timestamp
- Account transactions

03

Machine Learning Model and Results

Banking Domain

Machine Learning Model and Results

The selected model to predict the fraud was a k-means clustering algorithm. Several algorithms were test, given the nature of the data the k-means gabe the better results. The parameters choosen were:

- N_clusters = 10
- N_init = 20
- Algorithm = elkan

The model was tested with the silhouette score. In total the metric gave 0.31. Scale 0.0 – 1.0 where 1 would be possible if all the points lie in the same location.

04

Customer Descriptions

Banking Domain

Customer Description

Cluster Descriptions

	v1	v2	v4	v5	v6	day	hour	minute	second	v7	cluster	count
0	332.010602	22.031406	760.481687	23.045382	0.003773	5.472289	14.157430	28.758233	0.0	428.471084	0.0	1245
1	36.673154	19.704832	462.402550	29.877315	0.004247	8.742282	7.417450	27.688591	59.0	425.729396	1.0	745
2	83.127425	20.399254	517.286847	25.386660	0.003748	7.928105	20.183007	47.470588	59.0	434.159422	2.0	1072
3	32.529238	19.666025	460.169874	30.278756	0.004294	8.868246	5.105107	28.678756	0.0	427.640637	3.0	1351
4	205.195556	21.461222	647.565889	22.460222	0.003557	5.646667	16.453333	29.720000	59.0	442.370333	4.0	900
5	48.677220	20.394402	452.743436	20.672201	0.003065	6.084942	4.202703	29.635135	59.0	404.066216	5.0	518
6	9.367405	20.173497	451.051424	23.790902	0.003461	6.719937	19.590981	30.474684	0.0	441.684019	6.0	1264
7	255.618928	21.207035	1247.318425	34.920436	0.005460	9.023451	16.252931	29.323283	0.0	991.699497	7.0	597
8	41.198958	20.397135	450.981858	20.663976	0.003065	6.091146	4.196181	28.701389	0.0	409.782899	8.0	1152
9	303.354515	21.404348	1396.705686	35.218729	0.005572	9.009174	15.284404	30.013761	59.0	1093.351171	9.0	299

Customer Description

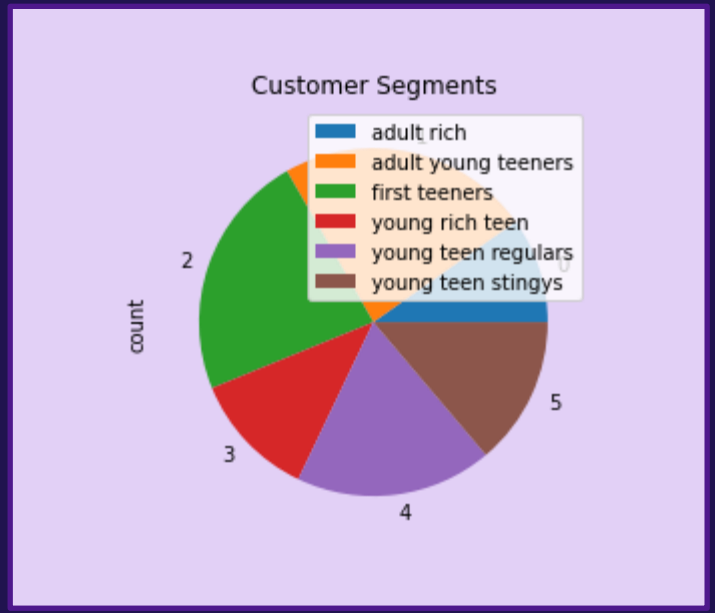
From the previous summarization description, we can infer the next situation group:

- 0 - 4 adult young teeners, they move the same amount of money as the adult and hold the best savings after the rich.
- 1 - 3 first teeners. People starting in their account, youngest and regulars.
- 2 are the young rich teen people, they do spend but hold the most savings in the 20's.
- 5 - 8 young teen regulars 20's. I do spend and save some.
- 6 young teen stingys 20's. they are the ones who move the less money and always save.
- 7 - 9 adult ~ 35 rich people They move money and save the most. Their balance in account is the highest.

Customer Description

Marketing Strategy

- You can start introducing pension plans to the group "adult young teeners".
- You can introduce special credit cards with reward programs or car loans to the "young rich teen people". They can pay the most.
- "First teeners" are not the option to introduce credit cards.
- You can offer car loans to "young teen regulars".
- The "young stingys" will not get any loans nor credit cards.
- The "adults" will most likely to accept pension deals, insurance deals and school bundle offers with credit cards.



Next Steps

This demo was a demonstration of what you can do with data available in a bank.

You can use more sophisticated techniques such as self organizing maps to find better clusters.

You can work with geo locations to create special offers base on geo location

You can improve your fraud detection by working with the proper amount of data

You can test more algorithms

Add external resources to the dataset

And many more ….

Thanks!

Benazir de la Rosa
Data Science Consultant