Laporan Ujian Akhir Semester Data Mining Penambangan Data



Oleh: 1301164163 Dhamir Raniah Kiasati Desrul

Program Studi Sarjana Informatika Fakultas Informatika Universitas Telkom Bandung 2019

A. Latar Belakang Masalah

Fintech adalah singkatan dari *Financial Technology* yang merupakan sebuah inovasi di dalam bidang jasa keuangan. Jasa keuangan yang dipegang oleh Fintech ini bergerak untuk usaha rakyat. Fintech diperuntukkan untuk membantu kalangan masyarakat berpenghasilan rendah yang ingin membuat suatu usaha pada bidang teknologi. Bantuan tersebut yakni menyediakan modal untuk meningkatkan pendapatan, menurunkan tingkat kemiskinan, dan menciptakan kesejahteraan. Fintech sudah membuka 100 cabang yang menyebar dari Banten hingg Banyuwangi.

Setiap cabang memiliki bermacam-macam indikator yang berbeda untuk dievaluasi untuk mengukur performa cabang. Data yang dikumpulkan untuk dievaluasi sejak tanggal pencairan pertama tahun 2016 hingga Oktober 2018.

Fintech akan bergantung dengan penggunaan framework evaluasi mengukur performa cabang, untuk itu diperlukan pendekatan metode-metode yang dalam data mining untuk dapat memberikan saran untuk merekomendasikan dengan tepat agar fintech dapat memiliki performa lebih baik.

B. Tujuan

Berikut merupakan tujuan dari framework yang dibuat:

- 1. Membantu untuk mengambil keputusan berdasarkan data yang diberikan
- 2. Mengatasi dan menganalisisi masalah yang terjadi pada setiap cabang
- 3. Mendapatkan informasi tentang data di *Fintech*.

C. Deskripsi Data

Data yang tersedia terdiri atas 3703 baris data beserta 11 kolom atribut. Data tersebut diperoleh dari kurung waktu 2016 sampai Oktober 2018. Berikut merupakan deskripsi atribut-atribut yang tersedia pada data:

Tabel 1.Deskripsi Atribut

		Jumlah values
Atribut	Keterangan Atribut	(kategori) di atribut
	cabang di Fintech,	
branch	umumnya di level	
	kecamatan	99

cutoff_date	tanggal batas	109
	area yang mencakup	
aroa	cabang-cabang di	
area	bawahnya, umumnya di	
	level kabupaten	31
	region yang mencakup	
region	area-area di bawahnya,	
	level provinsi	5
	tanggal pencairan	
first_date_disbursement	pinjaman pertama di	
	setiap cabang	44
	jumlah mitra yang	
active_borrowers	sedang memiliki	
active_bollowers	pinjaman di setiap	
	cabang	1713
active_agent	jumlah agen yang aktif	
active_agent	bekerja di setiap cabang	13
	persentase pinjaman	
deliquency_rate	bermasalah di setiap	
	cabang.	6
	sisa pinjaman yang	
outstanding	belum terbayar oleh	
	mitra	150
	jumlah pencairan	
weekly_disbursement	(rupiah) setiap minggu di	
	setiap cabang	128
 weekly_new_borrower_per_bp	jumlah mitra baru setiap	
	minggu di setiap cabang	56

D. Praproses

Praproses adalah proses untuk membersihkan atau memperbaiki data apabila ditemukan missing value, atau NaN di dalam dataset. Proses ini membantu untuk membuat data agar konsisten dan menghindari ketidakseimbanagan sebuah data.

Pada tugas ini akan dilakukan praproses dengan tahapan yakni mengganti nilai *missing value* dan NaN pada dataset, mengubah bentuk data yang *string* menjadi numerik pada data, pengambilan data yang akan diproses pada tahap selanjutnya.

A. Mengubah Missing Value

Data yang diperoleh masih memiliki missing value dan '-' pada kolom weekly_disbursement dan weekly_new_borrower_per_bp.

Berikut merupakan penjelasan jumlah data yang memiliki missing value dan '- pada data *Fintech:*

Out[31]:	branch	0
	cutoff_date	0
	area	0
	region	0
	first_date_disbursement	0
	active_borrowers	0
	active_agent	0
	deliquency_rate	0
	outstanding	0
	weekly_disbursement	6
	<pre>weekly_new_borrower_per_bp dtype: int64</pre>	165

Gambar 1. Data Missing Value

Berikut merupakan data fintech yang memiliki *missing value* seperti dibawah berikut:

Out[36]:									
	area	region	first_date_disbursement	active_borrowers	active_agent	deliquency_rate	outstanding	weekly_disbursement	weekly_new_borrower_per_bp
	AREAID_27	REGION_JR1	2017-07-22	988	4	>10%	2100	NaN	NaN
	AREAID_8	REGIONID_JT	2017-03-25	2586	11	1%-2.5%	6150	450.0	NaN
	AREAID_17	REGION_JH	2018-08-18	146	2	0%	350	NaN	1.0
	AREAID_20	REGION_JH	2017-06-03	1156	8	5%-10%	1850	NaN	1.0
	AREAID_21	REGION_JR1	2017-09-09	1743	8	1%-2.5%	3500	95.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	1501	8	>10%	3200	NaN	NaN
	AREAID_27	REGION_JR1	2017-08-19	1042	4	5%-10%	1700	NaN	NaN
	AREAID_20	REGION_JH	2017-06-03	659	8	5%-10%	900	NaN	NaN
	AREAID_20	REGION_JH	2017-06-03	676	5	5%-10%	950	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	991	4	>10%	2200	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	1512	5	>10%	3300	0.0	NaN
	AREAID_27	REGION_JR1	2017-08-19	1067	5	5%-10%	1850	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	996	3	>10%	2250	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	1527	5	>10%	3400	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	999	3	>10%	2350	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	1534	5	>10%	3550	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	1543	6	>10%	3650	0.0	NaN
	AREAID_29	REGION_JH	2017-09-09	1873	8	>10%	3900	0.0	NaN
	AREAID_27	REGION_JR1	2017-07-22	1001	3	>10%	2450	0.0	NaN

Gambar 2. Dataset Fintech

Pada Gambar 1 menjelaskan bahwa terdapat nilai NaN pada kolom atribut weekly_disbursement dan weekly_new_borrower. Untuk menghindari adanya NaN pada data maka akan dilakukan proses untuk mengganti NaN dengan nilai rata-rata (mean) dari nilai pada masing-masing kolom yang memiliki NaN value berdasarkan branch.

branch	cutoff_date	area	region	first_date_disbursement	active_borrowers	active_agent	deliquency_rate	outstanding	weekly_disbursement	w€
POINTID_33	2018-10-20	AREAID_27	REGION_JR1	2017-07-22	988	4	>10%	2100	NaN	
POINTID_16	2018-10-20	AREAID_8	REGIONID_JT	2017-03-25	2586	11	1%-2.5%	6150	450.0	

Gambar 3. Contoh data NaN

Seperti Gambar 2 pada kolom weekly_disbursement indeks ke-30 merupakan data NaN dan data tersebut terletak pada branch POINTID_33 yang dimana nilai yang diganti untuk nilai NaN tersebut bergantung pada data atribut weekly_disbursement pada branch POINTID_33 saja.

	branch	cutoff_date	area	region	first_date_disbursement	active_borrowers	active_agent	deliquency_rate	outstanding	weekly_disbursement
30	POINTID_33	2018-10-20	AREAID_27	REGION_JR1	2017-07-22	988	4	>10%	2100	NaN
155	POINTID_33	2018-10-13	AREAID_27	REGION_JR1	2017-07-22	991	4	>10%	2200	0.0
234	POINTID_33	2018-10-06	AREAID_27	REGION_JR1	2017-07-22	996	3	>10%	2250	0.0
378	POINTID_33	2018-09-29	AREAID_27	REGION_JR1	2017-07-22	999	3	>10%	2350	0.0
475	POINTID_33	2018-09-22	AREAID_27	REGION_JR1	2017-07-22	1001	3	>10%	2450	0.0
522	POINTID_33	2018-09-15	AREAID_27	REGION_JR1	2017-07-22	1007	3	>10%	2500	0.0
616	POINTID_33	2018-09-08	AREAID_27	REGION_JR1	2017-07-22	1010	3	5%-10%	2600	25.0
694	POINTID_33	2018-09-01	AREAID_27	REGION_JR1	2017-07-22	1004	3	2.5%-5%	2650	0.0
836	POINTID_33	2018-08-25	AREAID_27	REGION_JR1	2017-07-22	1010	3	1%-2.5%	2750	25.0
901	POINTID_33	2018-08-18	AREAID_27	REGION_JR1	2017-07-22	1006	3	0%-1%	2800	90.0
980	POINTID_33	2018-08-11	AREAID_27	REGION_JR1	2017-07-22	988	3	0%	2800	40.0

Gambar 4. Contoh data NaN (cont)

Untuk mengganti nilai NaN tersebut yang diperlukan adalah menghitung mean dari keseluruhan data pada branch POINTID_33, hasilnya yakni adalah 94.

:												
_		branch	cutoff_date	area	region	first_date_disbursement	active_borrowers	active_agent	deliquency_rate	outstanding	$weekly_disbursement$	w
	30	POINTID_33	2018-10-20	AREAID_27	REGION_JR1	2017-07-22	988	4	>10%	2100	94	
	155	POINTID_33	2018-10-13	AREAID_27	REGION_JR1	2017-07-22	991	4	>10%	2200	0	
	234	POINTID_33	2018-10-06	AREAID_27	REGION_JR1	2017-07-22	996	3	>10%	2250	0	
	378	POINTID_33	2018-09-29	AREAID_27	REGION_JR1	2017-07-22	999	3	>10%	2350	0	
	475	POINTID_33	2018-09-22	AREAID_27	REGION_JR1	2017-07-22	1001	3	>10%	2450	0	
	522	POINTID_33	2018-09-15	AREAID_27	REGION_JR1	2017-07-22	1007	3	>10%	2500	0	
	616	POINTID_33	2018-09-08	AREAID_27	REGION_JR1	2017-07-22	1010	3	5%-10%	2600	25	
	694	POINTID_33	2018-09-01	AREAID_27	REGION_JR1	2017-07-22	1004	3	2.5%-5%	2650	0	
	836	POINTID_33	2018-08-25	AREAID_27	REGION_JR1	2017-07-22	1010	3	1%-2.5%	2750	25	
	901	POINTID_33	2018-08-18	AREAID_27	REGION_JR1	2017-07-22	1006	3	0%-1%	2800	90	
	980	POINTID_33	2018-08-11	AREAID_27	REGION_JR1	2017-07-22	988	3	0%	2800	40	

Gambar 5. . Contoh data NaN (cont)

Proses itu terus dilakukan terhadap value NaN pada data sampai tidak ada lagi data yang bernilai missing value.

B. Mengubah Bentuk Data yang String menjadi Numerik pada Data

Perlu adanya mengubah bentuk data string menjadi data numerik agar mempermudah proses dan mempengaruhi proses performansi pada suatu metode yang akan digunakan. Pada proses ini data yang akan dirubah adalah data yang berlum berbentuk angka akan dirubah.

```
branch = labelencoder.fit_transform(dataset_fill_value['branch'])
cutoff_date = labelencoder.fit_transform(dataset_fill_value['cutoff_date'])
area = labelencoder.fit_transform(dataset_fill_value['area'])
region = labelencoder.fit_transform(dataset_fill_value['region'])
first_date_disbursement = labelencoder.fit_transform(dataset_fill_value['first_date_disbursement'])
deliquency_rate = labelencoder.fit_transform(dataset_fill_value['deliquency_rate'])
```

Branch, cutoff_date, area, region, first_date_disbursement, dan deliquency_rate adalah data yang berbentuk string maka akan diubah menjadi numerik. Data tersebut akan menjadi kategorik numerik. Bentuk kategori yang direpresentasikan dengan bentuk angka seperti gambar dibawah ini.

	branch	cutoff_date	area	region	first_date_disbursement	active_borrowers	active_agent	deliquency_rate	outstanding	weekly_disbursement	weekly_new_borrower_per_bp
0	41	107	19	3	20	1108	5	2	2150	20	3
1	34	107	0	2	15	1515	7	4	2500	110	1
2	25	107	12	2	12	1092	5	3	1850	75	7
3	61	107	4	4	26	683	3	0	1600	145	17
4	44	107	13	3	21	2294	8	1	4300	50	7
5	12	107	3	3	42	54	2	0	150	20	4
6	4	107	9	4	38	243	2	0	650	90	15
7	97	107	1	0	33	430	2	0	1150	110	19
8	79	107	18	1	34	235	2	0	600	85	13
9	93	107	15	4	35	516	3	0	1350	195	22
10	2	107	6	3	36	452	3	0	1200	140	16

Gambar 6. . Contoh data NaN (cont)

E. Analisis Pemilihan Algoritma

Data *Fintech* belum memiliki label atau kelas yang mendeskripsikan data setiap row dimana itu yang mendefinisikan bahwa data ini termasuk ke dalam *unsupervised learning*. Clustering merupakan salah satu pendekatan metode dari unsupervised learning dimana metode ini akan membantu untuk mengelompokkan data mana saja yang memiliki jarak yang dekat. Clustering yang digunakan adalah K-Means. K-Means adalah salah satu algoritma yang sangat tepat karena bentuk data yang diperoleh yakni numerik (berdasarkan hasil prampemrosesan yang telah dilakukan) dengan atribut yang sangat banyak dan beragam.

Pemilihan nilai K yang optimal akan disesuaikan dengan hasil dari metode elbow dan hasil performansi akan diukur menggunakan silhouette score.

F. Analisis Penentuan Parameter

Berikut merupakan parameter yang digunakan untuk menyelesaiakan permasalahan data *Fintech*:

- 1. Memilih data berdasarkan penyelesaian permasalahan pinjaman
- 2. Memilih atribut yang digunakan berdasarkan nilai kolerasi yang besar antar atribut
- 3. Mencari nilai K terbaik menggunakan Elbow Method dari rentan nilai 1-10
- 4. Clustering Data
- 5. Melakukan evaluasi pada performansi k-means menggunakan silhouette score

G. Hasil Percobaan

a. Memilih Data berdasarkan deliquence_rate dan cutoff_date

Memilih data berdasarkan penyelesaian permasalahan pinjaman

Pemilihan data dilakukan menyesuaikan dengan objective dari kasus pada data

Fintech. Atribut yang mendukung yakni deliquence rate dimana deliquence rate

ini merupakan atribut yang merepresentasikan pinjaman bermasalah. Ketika 0%

maka dinyatakan bahwa tidak terjadi peminjaman bermasalah. Karena pada

kasus ini hasil yang diperoleh untuk membantu menyelesaikan permasalahan

pinjaman maka data akan difilter yakni yang memiliki deliquence_rate > 0%.

Data berhasil di filter berdasarkan deliqunce_rate tapi perlu adanya evaluasi tiap

tahun dari setiap branch untuk mengetahui permasalahan peminjaman.

Permasalahan peminjaman terbesar banyak terjadi pada tanggal batas yang

ditentukan pada Tahun 2018. Oleh sebab itu data akan diambil berdasarkan

deliquence_rate > 0% dan cutoff_date >= Tahun 2018. Data yang berhasil di

filter sebesar 662 data.

b. Memilih atribut yang digunakan berdasarkan nilai kolerasi yang besar antar atribut

Data yang telah dilakukan proses filter akan diproses ke dalam perhitungan kolerasi untuk dapat mengetahui seberapa dekat hubungan antara satu atribut dengan atribut lainnya. Gambar merepresentasikan hasil nilai kolerasi.

Tabel 2. Tabel Korelasi

	branch	cutoff_date	area	region	ate_disburs	ive_borrow	ctive_agen	liquency_ra	outstanding	ly_disburse	w_borrow
branch	1	0,170253	-0,37915	0,474961	0,596668	0,055974	-0,10902	-0,18579	-0,16972	-0,07849	0,1193
cutoff_date	0,170253	1	-0,14236	0,344645	0,32234	0,214214	0,025966	0,535922	0,005515	-0,30156	-0,17079
area	-0,37915	-0,14236	1	-0,73196	-0,64152	0,197723	0,442942	0,037998	0,363702	0,116186	-0,12801
region	0,474961	0,344645	-0,73196	1	0,871898	-0,19274	-0,4578	0,077981	-0,29719	-0,17466	0,18844
ate_disburs	0,596668	0,32234	-0,64152	0,871898	1	-0,13891	-0,40162	0,068615	-0,23116	-0,13451	0,181137
ive_borrow	0,055974	0,214214	0,197723	-0,19274	-0,13891	1	0,622794	0,030345	0,892491	0,223636	-0,11937
ctive_agen	-0,10902	0,025966	0,442942	-0,4578	-0,40162	0,622794	1	-0,10493	0,582478	0,23699	-0,08663
liquency_ra	-0,18579	0,535922	0,037998	0,077981	0,068615	0,030345	-0,10493	1	0,004664	-0,2887	-0,07581
outstanding	-0,16972	0,005515	0,363702	-0,29719	-0,23116	0,892491	0,582478	0,004664	1	0,319285	-0,07906
ly_disburse	-0,07849	-0,30156	0,116186	-0,17466	-0,13451	0,223636	0,23699	-0,2887	0,319285	1	0,163414
ew_borrow	0,1193	-0,17079	-0,12801	0,18844	0,181137	-0,11937	-0,08663	-0,07581	-0,07906	0,163414	1

Dapat dilihat dari tabel bahwa data outstanding terhadap active_borrowers sangat berkolerasi dengan nilai 0.89.

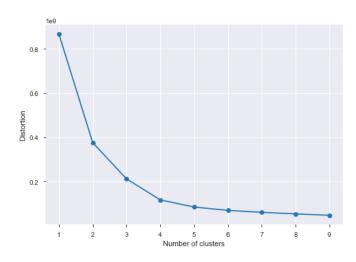
Tabel 3.Korelasi dengan Thereshold 0.5

	branch	cutoff_date	area	region	ate_disburs	ive_borrow	ective_agen	liquency_ra	outstanding	ly_disburse	weekly_n ew_borro wer_per_ bp
branch	TRUE										
cutoff_date	FALSE	TRUE									
area	FALSE	FALSE	TRUE								
region	FALSE	FALSE	FALSE	TRUE							
ate_disburs	TRUE	FALSE	FALSE	TRUE	TRUE						
ive_borrow	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE					
ctive_agen	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE				
liquency_ra	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE			
outstanding	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE		
dy_disburse	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	
ew_borrow	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE

Atribut yang akan dipakai menggunakan batasan nilai korelasi > 0.5 untuk dapat meminimalisir penggunaan atribut terlalu banyak yang dipakai dan dengan nilai 0.5 yang menyatakan bahwa kolerasi atribut seolah sudah separuhnya *similar*.

Atribut yang akan dipakai adalah branch, cutoff_date, area, region, first_date_disbursement, active_borrowers, deliquences_rate, out_standing.

c. Mencari nilai K terbaik menggunakan Elbow Method dari rentan nilai 1-10



Grafik 1. Metode Elbow

d. Clustering Data

```
Out[486]: 1 481
0 181
Name: clustering, dtype: int64
```

Gambar 7. Hasil data yang dicluster

e. Hasil Performansi Clusterring

Gambar 8. Hasil Performansi Clustering

H. Ringkasan Model Yang Diperoleh

Model yang dibangung merupakan hasil dari K-means yang dimana digunakan untuk menentukan cluster dari data. Pada kasus ini setelah dicoba menggunakan rentan angka 1-10 diperoleh K=2 sebagai nilai terbaik. Obeservasi ini berhasil membuat model yang dibangun dengan 2 cluster yakni *cluster* 0 dan *cluster* 1.

```
Out[490]: branch
                                       POINTID 9
          outstanding
                                            7800
                                        AREAID 8
          area
                                      REGION JR2
          region
          cutoff date
                                      2018-10-20
          active borrowers
                                            2816
          deliquency rate
                                            >10%
          first date disbursement
                                      2017-09-16
          clustering
          dtype: object
```

Gambar 9. Hasil Maksimum nilai Cluster 0

Pada Cluster 0 dapat dilihat bahwa Fintech mengalami peminjaman masalah terbesar pada tanggal 16 Juli 2017 untuk pencairan dana pertama kali dan batas tanggal 20 oktober 2018. Dari cluster 0 dapat diketahui bahwa peminjam yang aktif hingga batas tanggal sebanyak 2816.

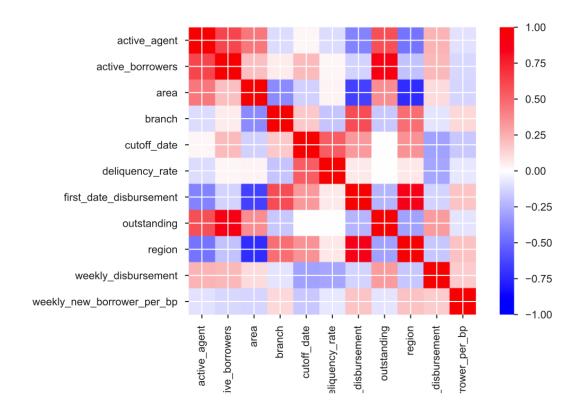
branch	POINTID_9
outstanding	3800
area	AREAID_8
region	REGION_JR1
cutoff_date	2018-10-20
active_borrowers	1962
deliquency_rate	>10%
first_date_disbursement	2018-01-13
clustering	1
dtype: object	

Gambar 10. Maksimum nilai Cluster 1

Pada Cluster 1 dapat dilihat bahwa Fintech mengalami peminjaman masalah terbesar pada tanggal 13 januari 2018 untuk pencairan dana pertama kali dan batas tanggal 20 oktober 2018. Dari cluster 1 dapat diketahui bahwa peminjam yang aktif hingga batas tanggal sebanyak 1962.

Dari kedua data tersebut dapat diketahui bahwa permasalahan juga terjadi pada area AREAID_8.

I. Interpretasi Model



Gambar 11. Representasi Korelasi

Kolerasi terdekat terjadi pada active_borrowers dan outstanding, terlihat dari warnanya yang hamper menyerupai dengan warna diagonal yakni merah tua yang merperesntasikan bahwa hubungannya bernilai hamper sama dengan koefisien 1.

	0%-1%	2.5%-5%	5%-10%	1%-2.5%	>10%
1	176	103	82	78	42
0	55	30	18	33	45

Gambar 12. Data berdasarkan delinquency_rate

	2018- 10-20				2018- 09-22								2018- 03-03							
1	24.0	24.0	24.0	23.0	21.0	20.0	18.0	18.0	18.0	18.0	 7.0	7.0	7.0	6.0	5.0	5.0	4.0	4.0	4.0	3.0
0	9.0	7.0	7.0	7.0	7.0	8.0	8.0	8.0	8.0	8.0	 1.0	1.0	1.0	NaN	1.0	NaN	NaN	NaN	NaN	NaN

Gambar 13. Cutoff Date

					2016- 09-17														2017- 04-22	
1	75.0	64.0	41.0	37.0	34.0	31.0	31.0	30.0	29.0	22.0	 16.0	11.0	10.0	7.0	5.0	1.0	1.0	NaN	NaN	NaN
0	22.0	NaN	NaN	NaN	7.0	NaN	NaN	NaN	NaN	NaN	 11.0	NaN	NaN	21.0	35.0	24.0	NaN	22.0	16.0	1.0

1. Model tersebut berisi 2 kelompok atau *cluster*:

- Cluster 1

Merupakan kelompok usaha atau mitra yang mengalami sedikit permasalahan peminjaman atau baik.

- Cluster 0

Merupakan kelompok usaha atau mitra yang mengalami besar permasalahan peminjaman atau buruk.

Dari 2 *cluster* tersebut presentase dari permasalahan peminjaman lebih sering terjadi pada cluster 0 dapat dilihat dari deliquency_rate, cutoff_date, first_date_disbursement.

Setelah ditelaah dengan banyaknya active_borrowers dan branch yang sedikit pada satu area akan mengakibatkan terjadinya permasalahan peminjaman, waktu batas yang akan lebih lama dan dana pencairan juga yang lebih lama bahkan jarang terjadi.Untuk saran sebaiknya dibuka cabang baru pada satu area tersebut untuk mencegah membludaknya peminjam atau mitra pada satu branch.