PORTOFOLIO DATA SCIENTIST

Analisis pada perusahaan Fintech

Laporan UAS Data Mining

Disusun Oleh :

Muhammad Fikri Septiawan 1301160789

Program Studi S1 Informatika

Fakultas Informatika

Telkom University

2019

A. Latar Belakang Masalah

Fintech atau *Financial Technology* adalah istilah yang digunakan untuk menyebut inovasi dalam bidang jasa keuangan atau finansial. Bisa juga diartikan dengan inovasi finansial yang diberi sentuhan teknologi modern. Atau merupakan segmen di dunia *start-up*. Tujuannya adalah membantu untuk memaksimalkan penggunaan teknologi untuk mengubah, mempertajam atau mempercepat berbagai aspek pelayanan keuangan. *Fintech* mempunyai salah satu proses bisnis yaitu meminjamkan modal kepada suatu usaha yang sudah berjalan maupun yang sedang dibangun.

Proses bisnis tersebut, dibutuhkan suatu pengambilan keputusan apakah keputusan untuk meminjamkan modal kepada suatu usaha tepat atau tidak (mendapatkan profit yang cukup besar maupun jangka waktu pengembalian modal). Masalah pengambilan keputusan tersebut mempunyai suatu solusi yaitu menggunakan *data minning* dengan mengacu pada suatu *dataset fintech*.

B. Tujuan

Adapun tujuan melakukan *data minning* ini sebagai berikut:

1. Untuk pengambilan keputusan peminjaman modal kepada suatu usaha,
2. Mendapatkan model yang tepat seperti syarat tertentu untuk dijadikan pengambilan keputusan pada *point* 1,
3. Mendapatkan informasi tentang perkembangan usaha yang telah diberikan modal oleh *fintech* ini, jika usahanya berkembang maka akan semakin mudah pencairan modal saat dibutuhkan oleh usaha tersebut.

C. Deskripsi Data

Dataset dengan nama data-uas.xlsx (bisa diunduh pada https://github.com/MFikriS/DataScientist/blob/master/Portofolio2/data-uas.xlsx) lalu harus di-*convert* menjadi data-uas.csv agar dapat diolah oleh bahasa pemrograman yaitu bahasa pemrograman python. Sesuai soal UAS digunakanlah baris data ke 500-1500 (1000 data), tidak mempunyai label dan memiliki 11 fitur (kolom) yaitu:

branch	cabang di Fintech, umumnya di level kecamatan.
cutoff_date	tanggal batas.
area	area yang mencakup cabang-cabang di bawahnya, umumnya di level kabupaten.
region	region yang mencakup area-area di bawahnya, level provinsi.
first_date_disbursement	tanggal pencairan pinjaman pertama di setiap cabang.
active_borrowers	jumlah mitra yang sedang memiliki pinjaman di setiap cabang.

active_agent	jumlah agen yang aktif bekerja di setiap cabang.
deliquency_rate	persentase pinjaman bermasalah di setiap cabang.
outstanding	sisa pinjaman yang belum terbayar oleh mitra
weekly_disbursement	jumlah pencairan (rupiah) setiap minggu di setiap cabang
weekly_new_borrower_per_bp	jumlah mitra baru setiap minggu di setiap cabang

Adapun tipe data dari fitur atau atribut dataset sebelum diubah:

```
In [200]: data.dtypes

Out[200]: branch                      object
          cutoff_date                 object
          area                        object
          region                      object
          first_date_disbursement     object
          active_borrowers            object
          active_agent                object
          deliquency_rate             object
          outstanding                 object
          weekly_disbursement         object
          weekly_new_borrower_per_bp  object
          dtype: object
```

Gambar 1. Tipe Data

Gambar 2. Error saat dilakukan proses perhitungan mean

Dilihat dari gambar 1, tipe data harusnya berupa *numerical* (*float* atau *int)* tetapi pada gambar tersebut masih tipe data *object atau string* (*error* untuk perhitungan seperti pada gambar 2) untuk fitur *active_borrowers, active_agent, deliquency_rate, outstanding, weekly_disbursement, weekly_new_borrower_per_bp* karena fitur tersebut berbentuk angka. Selain itu nilai dari fitur *weekly_new_borrower_per_bp* ada yang kosong (NaN) pada data ke 508, 521, 1485 seperti pada gambar 3 di bawah ini.

Gambar 3. Nilai kosong pada atribut weekly_new_borrower_per_bp

Dari 2 masalah tersebut (tipe data dan *missing value*), maka dataset belum berkualitas.

D. Praproses

Hasil codingan menggunakan bahasa pemrograman dapat diunduh pada https://github.com/MFikriS/DataScientist/blob/master/Portofolio2/uas.ipynb

1. Mengubah tipe data untuk fitur *active_borrowers, active_agent, deliquency_rate, outstanding, weekly_disbursement, weekly_new_borrower_per_bp* dari bertipe *object* menjadi *numerical*. Mengubah tipe data untuk , *cutoff_date, first_date_disbursement* menjadi date. Hasil pengubahan dapat dilihat pada gambar 4.

Gambar 4. Convert tipe data dan hasil convert

Hal tersebut dilakukan untuk proses perhitungan yang akan digunakan pada algoritma *data minning*.

2. Pemilihan baris data

   Pemilihan baris data dilakukan untuk pemilihan pola-pola data yang tepat. Melakukan proses tersebut karena setiap baris data memiliki pola-pola yang berbeda-beda. Pemilihan data dapat dilihat pada gambar 5.

Gambar 5. Pemilihan baris data

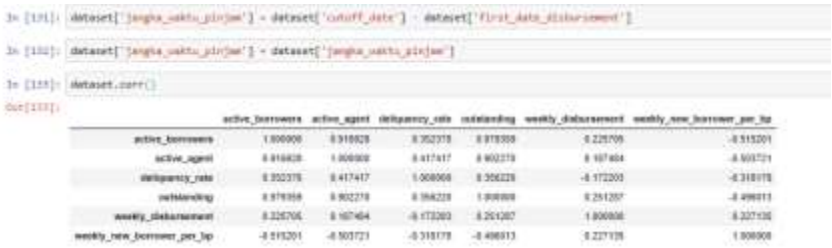
3. Mengubah tipe data fitur *deliquency_rate, region* menjadi tipe *categorical* karena fitur tersebut bertipe data *object* atau *string* (susah untuk diolah). Dilakukan untuk mengetahui tingkat pengaruh atribut dalam pengolahan data. Pra proses ini dapat dilihat pada gambar 6.

Gambar 6. Convert tipe data fitur deliquency_rate

4. Cek Missing Value pada atribut atau fitur dan mengisinya dengan teknik tertentu. Hal ini dilakukan untuk proses perhitungan agar menghasilkan perhitungan yang tepat.

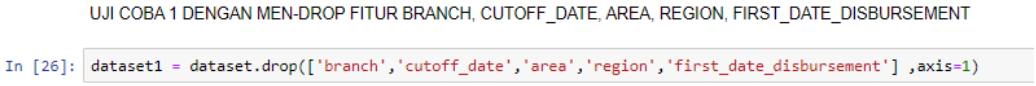
Gambar 7. Cek missing value dan pengisian missing value

5. Membuat kolom atau fitur baru yaitu jangka waktu pinjam karena fitur terserbut berpeluang untuk pengaruh pada pengolahan data. Fitur baru tersebut didapat dari pengurangan fitur *cutoff_date* dikurangi fitur *first_date_disbursement*. Pra proses tersebut dapat dilihat pada gambar 8.



Gambar 8. Pembuatan fitur baru yaitu jangka waktu pinjam

6. Melakukan drop fitur yang kurang berpengaruh pada pengolahan data. Hal ini dilakukan karena untuk meningkatkan nilai akurasi maupun *score*, dapat dilihat pada gambar 9.



Gambar 9. Drop fitur yang kurang berpengaruh

E. Analisis Pemilihan Algoritma

Berdasarkan kondisi dataset di atas yaitu tidak adanya label, maka metode yang tepat untuk dipilih adalah metode *clustering*. Metode *clustering* digunakan untuk mengelompokkan satu data ke data lain berdasarkan suatu atribut atau fitur. Metode *clustering* contoh algoritmanya yaitu *K-means*, *Aggromeratif Hierarcychal Clustering*, dan lainnya. Dari contoh algoritma *clustering* tersebut, dipilihlah algoritma *K-means* karena datanya mayoritas bertipe data *numerical* dan ngacak.

F. Analisis Penentuan Parameter

1. Pemilihan atribut atau fitur,
2. Pemilihan baris data (*row*),
3. Pemilihan metode klasifikasi atau *clustering*,
4. Pemilihan algoritmanya, jumlah *cluster*,
5. Nilai *silhouettte score*.

G. Hasil Percobaan

1. Pemilihan atribut atau fitur

   Fitur yang sangat berpengaruh pada pengolahan data *minning* disini adalah *active_borrowers, active_agent, deliquency_rate, outstanding, jangka_waktu_pinjam.*

Gambar 10. Drop fitur yang kurang berpengaruh dan yang tepat

2. Pemilihan baris data (*row*).
   Dipilih *row* data ke- 500-999 (500 data) karena mempengaruhi nilai *silhouette score*.

Gambar 11. Pemilihan baris data yang tepat

3. Pemilihan metode klasifikasi atau *clustering*.
   Menggunakan *clustering* karena dataset tidak mempunyai label, jika menggunakan klasifikasi maka harus ada label (*supervised*).

4. Pemilihan algoritma *clustering* dan jumlah *cluster*.
   Algoritma disini menggunakan *Agromerativ Hierarchical Clustering dan K-Means* dan jumlah cluster yang bagus yaitu sebanyak 2.

Gambar 12. Algoritma Kmeans

Gambar 13. Algoritma Agromerativ Hierarchical Clustering

5. Nilai *silhouettte score.*
   Sebagai penilaian kualitas dari *cluster*, semakin mendekati nilai 1 maka semakin bagus kualitas *cluster*.

Gambar 14. Algoritma Kmeans

Gambar 15. Algoritma Agromerativ Hierarchical Clustering

Berdasarkan gambar 14 dan gambar 15, algoritma *K-Means* lebih bagus daripada algoritma *Agromerativ Hierarchical Clustering* karena nilai *silhouette score K-Means* (0.6493175112255162) lebih besar dari nilai *silhouette score Agromerativ Hierarchical Clustering* (0.6385358408016295).

Dari hasil percobaan dengan analisis-analisis di atas, terdapat 2 kelompok atau *cluster* yaitu 0 dan 1. *Cluster* 0 berisi potensi *profit* besar bagi *fintech* dan Cluster 1 berisi potensi *profit* kecil bagi *fintech*. Dari *clustering* tersebut *fintech* dalam pengambilan keputusan dapat mengacu pada model *cluster* tersebut, Jika pemohon pinjaman merupakan *cluster* 0 maka *fintech* berani meminjamkan modal yang cukup besar dengan potensi *profit* cukup besar.

H. Ringkasan Model Yang Diperoleh

Dari model yang telah dibangun pada point G, didapatkan:
Terdapat 2 kelompok atau *cluster* yaitu 0 dan 1. *Cluster* 0 berisi potensi *profit* besar bagi *fintech* dan *Cluster* 1 berisi potensi *profit* kecil bagi *fintech* dan *Cluster* 1 berisi potensi *profit* kecil bagi *fintech*. Dari *clustering* tersebut *fintech* dalam pengambilan keputusan dapat mengacu pada model *cluster* tersebut, Jika pemohon pinjaman merupakan *cluster* 0 maka *fintech* berani meminjamkan modal yang cukup besar dengan potensi *profit* cukup besar juga.

I. Interpretasi Model

1. Import library numpy, pandas, dan KMeans

```
In [197]: import numpy as np
          import pandas as pd
          from sklearn.cluster import KMeans
```

2. Drop fitur yang kurang berpengaruh dalam proses data *minning*, yaitu *branch, cutoff_date, area, region, first_date,disbursement, weekly_new_borrower_per_bp, weekly_disbursement*

```
In [154]: dataset4 = dataset.drop(['branch','cutoff_date','area','region','first_date_disbursement','weekly_new_borrower_per_bp'] ,axis=1)
```

```
In [156]: dataset4 = dataset4.drop(['weekly_disbursement'] ,axis=1)
```

3. Pemilihan *row* atau baris data yang tepat yaitu baris ke-500-999

```
In [162]: dataset46 = dataset4.iloc[0:500,0:5]
```

4. Model Algoritma *K-Means Clustering* dengan jumlah 2 melakukan iterasi maksimum 10000 dan peniliaian menggunakan *silhouette score*

```
In [186]: kmeans = KMeans(n_clusters=2, max_iter=10000)
          dataset46['cluster'] = kmeans.fit_predict(dataset46)

          from sklearn.metrics import silhouette_score
          score = silhouette_score(dataset46, dataset46['cluster'], metric='euclidean')
          score
Out[186]: 0.6493175112255162
```

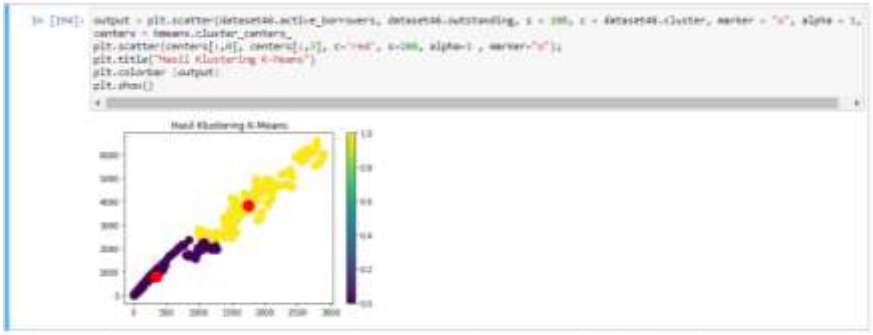
```
In [176]: kmeans = KMeans(n_clusters=3, max_iter=10000)
          dataset46['cluster'] = kmeans.fit_predict(dataset46)

          from sklearn.metrics import silhouette_score
          score = silhouette_score(dataset46, dataset46['cluster'], metric='euclidean')
          score
Out[176]: 0.6429690870560772
```

5. Hasil *cluster*

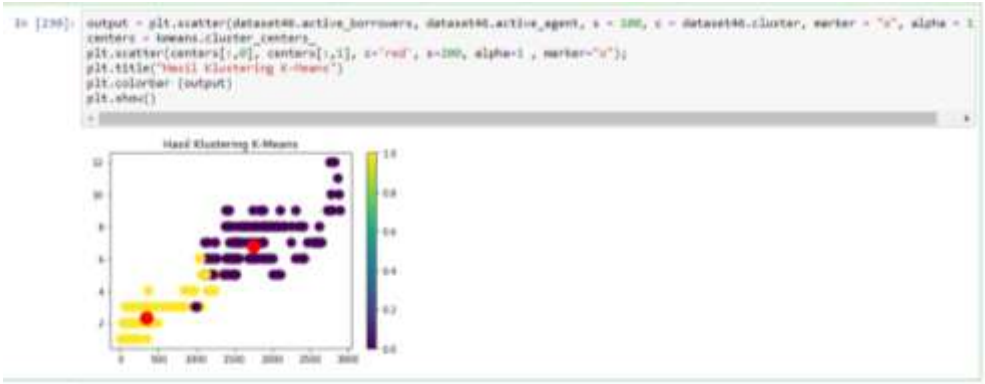
```
In [190]: dataset46['cluster'] = dataset46.cluster

In [191]: dataset46
```

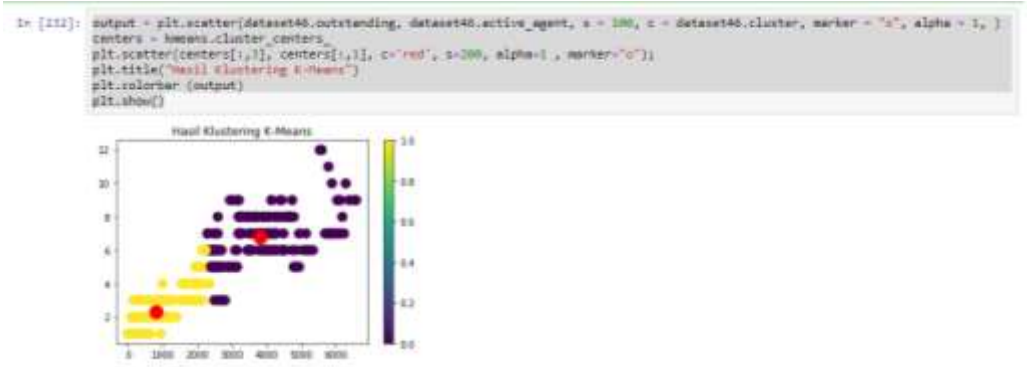
	active_borrowers	active_agent	delinquency_rate	outstanding	jangka_waktu_pinjam	cluster
500	1213.0	5.0	0	1500.0	200	0
501	416.0	3.0	0	1650.0	133	0
502	489.0	3.0	0	1280.0	112	0
503	425.0	3.0	0	1680.0	133	0
504	480.0	2.0	0	1300.0	159	0
505	375.0	2.0	0	1190.0	77	0
506	37.0	2.0	0	190.0	21	0
507	2700.0	12.0	0	5550.0	539	1
508	1600.0	7.0	5	4650.0	371	1
509	420.0	3.0	0	1650.0	112	0
510	145.0	2.0	0	490.0	35	0
511	114.0	1.0	0	380.0	42	0
512	1704.0	9.0	5	2950.0	497	1
513	1131.0	5.0	0	2490.0	371	1
514	160.0	2.0	0	520.0	56	0
515	280.0	2.0	0	830.0	42	0
516	1999.0	9.0	5	4290.0	539	1
517	2600.0	7.0	0	6250.0	413	1
518	1227.0	5.0	0	1950.0	266	0
519	791.0	3.0	0	3850.0	56	0

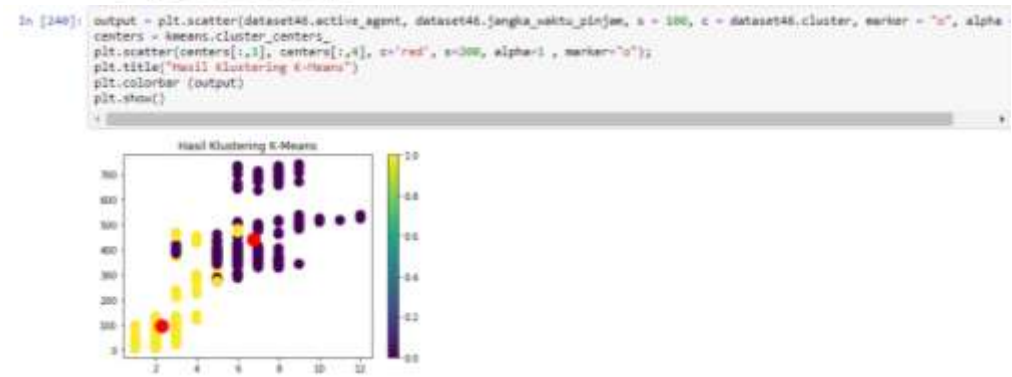
6. Visualisasi *cluster*



SUMBU X = FITUR/ATRIBUT 'active_borrowers' DAN SUMBU Y = FITUR/ATRIBUT 'outstanding' DIMANA BULAT MERAH ADALAH CENTROID







Model diatas menunjukan tujuan pada point B telah tercapai, karena:

1. Model tersebut berisi 2 kelompok atau *cluster*:
   - *Cluster* 0
   Merupakan potensi *profit* besar bagi *fintech*
   - *Cluster* 1
   Merupakan potensi *profit* kecil bagi *fintech*

Dari 2 *cluster* tersebut dapat digunakan pengambilan keputusan yaitu pemberian pinjaman modal dengan cukup banyak nominalnya kepada cluster 0 karena akan mendapat *profit* yang lebih besar juga daripada peminjaman cluster 1. Karena pada *cluster* 0 terdapat jangka_waktu_pinjam yang lama yaitu > 300 hari maka bunga yang didapat semakin tinggi, hal ini akan mendapatkan profit yang besar. Selain itu jumlah agen yang aktif bekerja >= 6 pada *cluster* 0 akan mengawasi pinjaman yang diberikan, hal ini akan meminimalkan resiko (kerugian) yang ada.

2. *Fintech* memberikan syarat pinjaman yaitu jika jumlah mitra yang sedang memiliki pinjaman di setiap cabang berjumlah sedikit (0-700), sisa pinjaman yang belum terbayar oleh mitra berjumlah sedikit (0-2100) juga, jumlah agen yang aktif bekerja di setiap cabang berjumlah banyak (>=6), jangka_waktu_pinjam > 300 hari (akan menghasilkan bunga yang besar makan akan mendapatkan profit yang besar juga), maka usaha atau mitra dapat meminjam lagi

3. Perkembangan usaha/mitra dapat dilihat pada model yang telah dibangun. Semakin berkembang usaha/mitra maka semakin patuh atau cepat dalam membayar pinjaman (dilihat dari jumlah mitra yang sedang memiliki pinjaman di setiap cabang yang sedikit dan sisa pinjaman yang belum terbayar oleh mitra berjumlah sedikit pada *cluster* 0). Hal ini akan memberikan *profit* untuk *fintech*.