Laporan Tugas 1

Penambangan Data dan Intelegensia Bisnis

Zaki Raihan

1606878505

Penambangan Data dan Intelegensia Bisnis - B

Program Studi Sistem Informasi

Fakultas Ilmu Komputer

Universitas Indonesia

Depok

2018

1. Pendahuluan

   Laporan ini terkait dengan langkah-langkah yang saya lakukan ketika bertemu dengan suatu data. Sesuai dengan yang diajarkan di kelas, maka saya akan melakukan data preparation yang digunakan agar data yang akan kita proses dalam kondisi optimal (tidak mengandung banyak bias) sehingga penggunaan data kedepannya pun akan menjadi lebih baik.

2. Deskripsi Data

   Data yang digunakan adalah data *home equity load*, yaitu sebuah pinjaman yang mana peminjam menggunakan *equity* dari rumahnya sebagai agunan dasarnya.
   Data ini memiliki beberapa variabel didalamnya yaitu sebanyak 13 variabel:
   . BAD: 1 = peminjam terkena gagal pinjaman atau mengalami tunggakan serius; 0 = peminjam membayar hutang
   . LOAN: banyaknya pinjaman yang ingin diminta
   . MORTDUE: Amount due on existing mortgage
   . VALUE: nilai dari properti saat ini
   . REASON: alasan melakukan peminjaman, DebtCon = debt consolidation; HomeImp = home improvement
   . JOB: jenis pekerjaan
   . YOJ: lama sudah bekerja di pekerjaan saat ini
   . DEROG: Number of major derogatory reports
   . DELINQ: Number of delinquent credit lines
   . CLAGE: Age of oldest credit line in months
   . NINQ: Number of recent credit inquiries
   . CLNO: Number of credit lines
   . DEBTINC: Debt-to-income ratio

3. Data Preparation

   Untuk melakukan persiapan data, saya menggunakan bahasa pemrograman python dengan bantuan beberapa modul seperti numpy dan pandas. Pandas sendiri sudah cukup sering digunakan di dunia pengolahan data. Dengan menggunakan modul panda ini, saya berusaha untuk membersihkan bagian data yang kurang berguna seperti kolom yang berlebih ataupun adanya outlier pada dataset yang diberikan. Untuk outlier sendiri, saya akan ganti dengan median dari kolom yang bersangkuta (tempat outlier berada). Saya merasa pergantian dengan median lebih baik daripada dengan menggunakan mean, karena mean sendiri juga dipengaruhi oleh keberadaan dari outlier itu sendiri. Saya berusaha untuk mengubah data yang saya rasa terjadi kesalahan pengetikan, sehingga data lebih baik. Saya melakukan normalisasi data dengan menggunakan metode minmax agar data yang ada hanya pada rentang nilai 0-1 sehingga tidak terlihat terlalu jauh perbedaan antar kolom data. Untuk data yang bersifat katagorikal, maka akan saya ubah menjadi numerik, hal itu saya lakukan agar pada saat data diolah, maka kita dapat melihat korelasi dari setiap variabel data.

4. Hasil/Temuan
   a. Ada data yang memiliki kolom lebih dari 13, sehingga kolom-kolom yang melebihi dari 13 akan saya hapus.
   b. Adanya kesalahan penulisan pada variabel “JOB”, data tersebut akhirnya saya anggap kosong dan saya ganti dengan nilai 0, (nilai nol merupakan hasil mapping dari nilai Other pada kolom JOB)

c. Terdapat cukup banyak outlier pada data (pada kolom tertentu) sehingga saya mengubah outlier tersebut menjadi nilai median dari data yang bersangkutan.
d. Nilai dari data pada tiap kolom cukup bervariasi mulai dari satuan, puluhan, bahkan ada variabel yang mengandung data dengan nilai ratusan ribu. Untuk itu saya melakukan normalisasi dengan metode minmax sehingga rentang nilai di setiap variabel data sama, yaitu 0 s.d. 1.
e. Terdapat korelasi yang cukup baik antar data dimana:
    i. Data variabel "BAD" memiliki korelasi yang cukup baik secara negatif dengan variabel "LOAN", "VALUE" dan "CLAGE" serta korelasi yang cukup baik secara positif dengan variabel "DEROG", "DELINQ", dan "DEBTINC".
    ii. Data variabel "VALUE" memiliki korelasi yang cukup baik dengan beberapa variabel lain seperti "LOAN", "MORTDUE", "JOB", dan "CLNO"

Untuk korelasi antar variabel bisa dilihat dari gambar dibawah ini:

	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
BAD	1.0	-0.15	-0.093	-0.11	-0.068	-0.022	-0.064	0.22	0.33	-0.15	0.064	-0.042	0.12
LOAN	-0.15	1.0	0.16	0.33	0.33	0.064	0.12	-0.014	-0.061	0.095	0.028	0.12	0.083
MORTDUE	-0.093	0.16	1.0	0.77	0.061	0.24	-0.075	-0.029	-0.018	0.091	-0.02	0.28	0.12
VALUE	-0.11	0.33	0.77	1.0	0.11	0.32	0.0044	-0.05	-0.0021	0.19	-0.013	0.35	0.12
REASON	-0.068	0.33	0.061	0.11	1.0	-0.019	-0.053	-0.0027	-0.027	-0.01	0.062	0.14	0.046
JOB	-0.022	0.064	0.24	0.32	-0.019	1.0	-0.012	-0.037	0.024	0.088	-0.036	0.24	-0.088
YOJ	-0.064	0.12	-0.075	0.0044	-0.053	-0.012	1.0	-0.059	0.054	0.22	-0.031	0.042	-0.079
DEROG	0.22	-0.014	-0.029	-0.05	-0.0027	-0.037	-0.059	1.0	0.15	-0.077	0.39	0.011	0.025
DELINQ	0.33	-0.061	-0.018	-0.0021	-0.027	0.024	0.054	0.15	1.0	0.013	0.015	0.13	0.043
CLAGE	-0.15	0.095	0.091	0.19	-0.01	0.088	0.22	-0.077	0.013	1.0	-0.082	0.18	-0.032
NINQ	0.064	0.028	-0.02	-0.013	0.062	-0.036	-0.031	0.39	0.015	-0.082	1.0	0.0063	0.079
CLNO	-0.042	0.12	0.28	0.35	0.14	0.24	0.042	0.011	0.13	0.18	0.0063	1.0	0.13
DEBTINC	0.12	0.083	0.12	0.12	0.046	-0.088	-0.079	0.025	0.043	-0.032	0.079	0.13	1.0

f. Untuk metode pengolahan data dapat dilihat dalam lampiran

5. Kesimpulan

Data Preparation merupakan proses yang cukup penting dalam pengolahan suatu data. Hal tersebut dikarenakan sering kali data yang kita peroleh terdapat banyak kekurangan seperti adanya data yang hilang, kesalahan penulisan, sampai adanya kolom yang berlebih. Selain itu Data preparation juga penting untuk mengurangi bias yang dimiliki oleh data yang kita miliki. Sering kali ada saja data yang nilainya jauh berbeda dari data yang lain (Outlier) sehingga dapat mengganggu kita pada saat melakukan pengelolaan data. Jika Data Preparation dilakukan dengan baik, maka hasil yang kita dapatkan pada saat pengelolaan data nantinya juga akan baik pula.

Untuk dokumentasi source code dari data preparation yang saya lakukan dapat dilihat pada:

https://github.com/zakiraihan/PDIB-ZakiRaihan-1606878505/tree/master/Tugas%201