Universitas Indonesia

Tugas 1: DATA PREPARATION

PENAMBANGAN DATA DAN INTELEGENSIA BISNIS

BAHY HELMI HARTOYO PUTRA
1606918124

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI SISTEM INFORMASI
DEPOK
OKTOBER 2019

Pendahuluan

Data preparation merupakan hal yang penting untuk dilakukan sebelum memulai sebuah *data science*/*machine learning project*. Terkadang kita berfikir bahwa Data Scientist banyak menghabiskan waktunya pada proses perancangan algoritma *machine learning* (*modelling*). Namun realitanya, kebanyakan Data Scientist menghabiskan waktunya untuk menjalankan proses *data preparation*.

Tugas ini akan mencakup proses awal dari sebuah *machine learning/data science project*, yaitu *data preparation. Data preparation* akan dilaksanakan menggunakan beberapa teknik yang telah dipelajari seperti eksplorasi, transformasi, *smoothing, outlier analysis,* dan *imputation*.

Deskripsi Data

Sebelum melakukan operasi pada sebuah data, sangat penting bagi seorang Data Scientist untuk dapat memahami data yang dimiliki. Selain itu, pengetahuan tentang apa yang ingin dicapai menggunakan data tersebut juga harus dimilki. Pemahaman akan dua hal ini mempemiudah seorang Data Scientist dalam melakukan pengambilan keputusan saat menjalankan proses *data preparation*.

Data yang akan diolah adalah *data-t1.csv* yang tersedia pada deskripsi Tugas 1. Berikut merupakan deskripsi mengenai data tersebut:

Nama Variabel	Tipe Data	Penjelasan
loan_status	(str, categorical)	Variabel degan *multiple levels* yang menandakan status pinjaman seseorang (e.g. Charged off, Current, Default, Fully Paid, etc)
loan_amnt	(int, discrete)	Variabel yang berisikan jumlah pinjaman yang dimilki seseorang (e.g. 1000, 2000, 3000, etc.)
int_rate	(float, continuous)	Variabel yang berisikan bunga pinjaman seseorang (e.g. 13.56, 2.5, 12.3, etc.)
grade	(str, categorical)	Variabel yang berisikan tingkat *employment* seseorang (e.g. A, B, C, etc.)
emp_length	(str, categorical)	Variabel yang berisikan

		durasi *employment* seseorang (e.g. 4 years, 5 years, etc.)
home_ownership	(str, categorical)	Variabel yang berisikan status kepemilikan rumah seseorang (e.g. RENT, ANY, etc.)
annual_inc	(float, continuous)	Variabel yang berisikan total penghasilan tahunan seseorang (e.g. 111.24, 123.14, etc.)
term	(str, categorical)	Variabel yang berisikan term seseorang (e.g. 60 months, 36 months)

Data awal berisikan 149997 baris dan 8 kolom.

Data Preparation

Data preparation merupakan sebuah proses untuk melakukan persiapan terhadap sebuah *raw data*. Raw data yang baru di*mining*/diterima Dalam melakukan *data preparation* kali ini, akan ada beberapa tahapan yang dilakukan, yaitu sebagai berikut:

- **Load Data**: melakukan *load* terhadap data yang ingin diolah, baik bersumber sebuah *cloud storage* (GCS/BigQuery, AWS) ataupun pada kasus ini dari data *local* yaitu sebuah file *csv*.
- **Eksploration & Data Checking**: pada tahapan ini yang dilakukan adalah melakukan *sanity check* terhadap data dan melihat *errors* yant teradapat pada data. Errors dapat merupakan *miss-recorded* data ataupun *null values* data.
- **Data Smoothing:** terkadang perbedaan antara data yang minor tidak terlalu signifikan dan tidak menurunkan performa. Data *smoothing* dapat membuat data-data yang memiliki perbedaan tidak signifikan tersebut menjadi lebih seragam.
- **Data Imputation:** *imputation* dilakukan untuk melakukan *replacement* terhadap *null values* yang ada pada suatu kolom. Teknik *imputation* ada berbagai macam, dapat *berupa mean/median/mode imputation*.
- **Data Transformation:** data terkadang perlu untuk dilakukan transformasi. Hal ini membuat sebuah data menjadi lebih *usable*, lebih mudah diinterpretasikan, serta *useful* untuk digunakan pada tahapan berikutnya.
- **Outliers Treatment:** *outliers* dapat membuat data salah diinterpretasikan saat melihat statistik dari data tersebut. Sekumpulan atau sebuah *outlier* dapat menarik *mean* ke arah negatif/positif yang mana sebenarnya pesebaran distribusi data tidak berada di titik tersebut. Outliers perlu ditangani dengan cara dihilangkan.

Proses *data preparation* dilakukan menggunakan bantuan bahasa pemrograman Python pada sebuah tools, yaitu Jupyter Notebook. Library yang digunakan pada *data preparation* kali ini adalah NumPy, pandas, dan juga seaborn. Detil pengerjaan dan *notebook* dapat diakses pada lampiran.

Hasil/Temuan

- Terdapat **2 *bad lines*** pada *raw* data, yaitu pada line 52431 dan 131201. Seharusnya tabel ini hanya memiliki 8 kolom, namun pada baris tersebut data tersebar pada 9 kolom.
- Data *raw* tanpa *bad lines* memiliki **149997 baris** dan **8 kolom**.
- *Miss-recorded data*, di luar *null values*, terdapat pada kolom-kolom:
    - *loan_status* = 1213 baris
    - *emp_length* = 35 baris
    - *home_ownership* = 5 baris
- ***Null values** data* terdapat pada kolom *emp_length*, yaitu sebanyak **9%** dari data atau sebanyak **13668** baris.
- ***Smoothing*** perlu dilakukan pada kolom *annual_inc*, *smoothing* dilakukan pada data yang bertipe *float* pada kolom tersebut. Operasi ***round*** dilakukan untuk mengubah ke *nearest integer*.
- **Imputation** dilakukan pada kolom emp_length dimana terdapat 9% *null values*. Karena raw data bertipe *qualitative* (*categorical*), maka ***mode imputation*** dilakukan untuk mengisi *null values.*
- **Transformasi** dilakukan kepada kolom-kolom yang bertipe ordinal dan nominal.
    - Kolom yang bertipe **ordinal** (*grade* & *emp_length*) ditransform ke dalam numerical. Grade A, G melambangkan tingkatan *risk* yang dimiliki, A paling kecil, G paling besar, sehingga dapat di*transform* menjadi angka 1-7. Employment length dapat ditransform menjadi angka 0-10, dimana 0 melambangkan *employment length* dibawah 1 tahun, dan 10 merupakan lebih dari 10 tahun.
    - Kolom yang bertipe **nominal** (*loan_status, home_ownership, term*) ditransform dengan bantuan pandas.get_dummies untuk menjadikan sebagai kolom *numerical boolean* (1,0) dengan membuat *value* menjadi judul kolom.
- **Outliers treeatment** perlu dilakukan untuk kolom *annual_inc* dan *int_rate*. Treatment yang dilakukan adalah melakukan *outliers removement* berdasarkan threshold yang telah ditentukan. Penentuan *threshold* dilakukan berdasarkan aturan berikut:
    - Q1 - 1.5 IQR, Q3 + 1.5 IQR (diterapkan pada kolom *annual_inc*)
    - Mean +- 2 Standard Deviation (diterapkan pada kolom *int_rate*)
- Data final yang telah siap untuk digunakan pada proses berikutnya berisikan **128548 baris** dan **17 kolom.**

Kesimpulan

Dari proses *data preparation* yang dilakukan, ada beberapa poin-poin penting yang dapat disimpulkan, yaitu:

- Melakukan *data preparation* sebelum memulai sebuah *machine learning project* sangatlah penting. Data tidak dapat langsung digunakan karena data memiliki resiko mengandung *error/miss-recorded* data.
- Memahami teknik *data preparation* sangatlah penting, dengan pemahaman tersebut kita akan lebih mudah untuk melakukan operasi-operasi yang diinginkan pada data yang dimilki.
- Memahami konteks serta setiap variabel yang ada dalam data yang dimiliki juga penting. Tanpa mengetahui hal tersebut, validasi terhadap data akan sulit untuk dilakukan. Pemahaman ini juga akan membantu dalam proses *brainstorming* saat ingin memulai *data preparation*.

Lampiran

Notebook dan hasil dari data preparation dapat diakses pada link berikut:
http://bit.ly/Tugas1-PDIB-Bahy