

Stage 1



Kelompok: Car-a-thon

Stage: 1

Mentor: Stephanie

Pukul/ Tanggal: 21.00 WIB/ 13 Mei 2022

Pembagian tugas di stage ini:

Nama: Bagus Ganjar Lugina

Tugas: Descriptive Statistics, Univariate Analysis, Business Insight, Notulen Mentoring, Source Code, Laporan Project.

Nama: M. Harun Arrasyid

Tugas: Descriptive Statistics, Multivariate Analysis, Business Insight, Notulen Mentoring, Source Code, Materi PPT.

Nama: Bernadetha Stella

Tugas: Descriptive Statistics, Multivariate Analysis, Business Insight, Notulen Mentoring, Source Code, Laporan Project.

Nama: Samuella Magdalena E

Tugas: Descriptive Statistics, Multivariate Analysis, Business Insight, Notulen Mentoring, Source Code, Materi PPT, Laporan Project.

Nama: Raihan Kurniasugianto

Tugas: Descriptive Statistics, Latar Belakang Masalah, Univariate Analysis, Business Insight, Notulen Mentoring, Source Code, Laporan Project.

Nama: M Rifqi Sarosa

Tugas: Descriptive Statistics, Univariate Analysis, Business Insight, Notulen Mentoring, Source Code, Materi PPT.



Kelompok: Car-a-thon

Stage: 1

Mentor: Stephanie

Pukul/ Tanggal: 21.00 WIB/ 13 Mei 2022

Poin pembahasan:

1. Latar belakang

- Cari sumber eksternal untuk pendukung latar belakang, misalnya melihat kesalahan penilaian mobil berapa persen dan sefatal apa (apakah harga tinggi dijual rendah, berapa yang ga jadi beli karena harga terlalu tinggi, dsb.).
- *Appraisal time* tidak terlalu butuh *research*, tapi bisa dicari informasi tentang penilaian harga mobil biasanya memakan waktu berapa lama,
- Untuk latar belakang sebaiknya di-*support* dengan sumber eksternal terlebih dahulu, tetapi jika tidak ada bisa menggunakan asumsi.

2. Descriptive Statistics:

- Untuk *present* akhir perlu dibagi 2 bagian: *overview* dan kebersihan data (*numerical* dan *categorical*, berapa baris dan kolom, kebersihan data, *missing data* dan persentasenya serta tindak lanjut yang perlu dilakukan),
- Lebih ditekankan kembali untuk penggalian insightnya khususnya yang menjadi *highlight* saja ketika di materi PPT.

3. Univariate Analysis

- Data numerical bisa menggunakan *box plot*, *violin plot*, *histogram* (umumnya *box plot* atau *histogram* lebih mudah dipahami awam), dikarenakan ketika presentasi akhir nanti ada beberapa stakeholder *non-technical* (tim bisnis) sehingga perlu ditampilkan distribusi dengan grafik yang mudah dipahami,
- Melihat distribusi (dengan *histogram*), melihat *outliers* (dengan *boxplot*) dan apa yang mau dilakukan (*keep*, *cut* batas tertinggi dan terendah, dll).

4. Multivariate Analysis

- *Heatmap* untuk melihat korelasi linear. Untuk variabel numerik yang memiliki korelasi sangat tinggi (>0.95), kemungkinan redundant. Jika diambil sebagai fitur, maka dapat menyebabkan akurasi menurun → tinjau fitur mmr. Batas suatu fitur dapat dianggap redundant adalah sebesar >0.95 . Korelasi negatif maupun positif yang sangat tinggi dapat dianggap redundant,
- Korelasi kecil dari heatmap belum tentu tidak berkorelasi dan harus drop feature, untuk melihat redundant maka selanjutnya perlu dianalisis dengan korelasi linear juga,
- Data *categorical*/biasanya menggunakan KDE Plot untuk melihat distribusi. Cek persentase *low*, *medium*, *high* untuk setiap kategorinya,
- Dapat menggunakan *count plot* per kategori (untuk kategori yang banyak dapat dikelompokkan dulu). Bisa dibuat *threshold* untuk data categorical kurang lebih 10,
- Untuk data *categorical*/perlu melihat korelasi dari kde, tidak bisa secara angka dengan heatmap. Dengan menggunakan kde, korelasi untuk data *categorical*/dapat dilihat dari semakin terpisahnya masing-masing kategori (semakin terpisah-pisah maka semakin berpengaruh).



Kelompok: Car-a-thon

Stage: 1

Mentor: Stephanie

Pukul/ Tanggal: 21.00 WIB/ 13 Mei 2022

Hasil Diskusi:

- Sales revenue bisa diukur dengan menghitung rata-rata kenaikan harga mobil setelah menggunakan model *machine learning* yang dibuat dibantu dengan sumber eksternal (Ekspektasi dari tim rakamin dalam melihat *business metrics* harus memiliki korelasi antara model yang dibuat dengan *metrics*),
- Untuk KDE plot yang memiliki beberapa puncak dapat didefinisikan bahwa kategori tersebut multimodal sehingga mendekati distribusi normal, contohnya pada kolom *condition*,
- Untuk data *categorica*/dikelompokkan kembali mengacu pada banyaknya data dan hubungan antar fitur. Contoh: ambil topnya dan kelompokkan sisanya menjadi others. Intinya melihat hubungan dan tindak lanjut yang perlu dilakukan,
- Untuk melihat hubungan antar fitur, *selling price* bisa dibagi high dan low, kemudian melihat persentase tiap merk mobil di masing-masing kategori selling price. Dapat juga dengan melihat rata-rata harga setiap merk, kemudian dikelompokkan masing-masing merk yang memiliki range harga yang sama,
- Fitur *seller* dapat dikelompokkan dengan mengambil yang beberapa yang tertinggi saja (*top seller*) dan sisanya dibuat sebagai “others”, sedangkan fitur selling date bisa diambil bulan atau tahunnya saja.

Tindak Lanjut:

- Mencari sumber data eksternal atau membuat suatu asumsi untuk memberikan statement yang kuat terhadap business metrics **Sales Revenue** dan **Appraisal Time**.
- Menambahkan analisis yang menampilkan overview data (jumlah row dan kolom dan persentase missing values).
- Handle lowercase dan uppercase yang ada pada setiap kolom dengan tipe data string, untuk menyederhanakan data sebelum melakukan **Univariate Analysis**.
- Ubah Univariate Analysis dengan menggunakan **boxplot/histogram**.
- Menganalisis dan melihat korelasi fitur kategorikal dengan menggunakan KDE Plot
- Melakukan pengelompokan kolom yang memiliki kategori banyak berdasarkan **selling price (target)** untuk mempermudah dalam melihat distribusi. Contoh: Mengelompokkan top 10 make (merk) mobil terjual berdasarkan segmentasi harga low, med, high.

Stage 2



Kelompok: Car-a-thon

Stage: 2

Mentor: Stephanie

Pukul/ Tanggal: 21.00 WIB/ 20 Mei 2022

Pembagian tugas di stage ini:

Nama: Bagus Ganjar Lugina

Tugas: Data pre-processing (fitur: seller, mmr, sellingprice), Compile Coding, Laporan Project, Materi PPT.

Nama: Bernadetha Stella

Tugas: Data pre-processing (fitur: year, make, model), Laporan Project, Notulen Mentoring, Materi PPT.

Nama: Samuella Magdalena E

Tugas: Data pre-processing (fitur: vin, state, condition, saledate), Laporan Project, Notulen Mentoring, Materi PPT.

Nama: Raihan Kurniasugianto

Tugas: Data pre-processing (fitur: odometer, color, interior), Compile Coding, Laporan Project, Materi PPT.

Nama: M Rifqi Sarosa

Tugas: Data pre-processing (fitur: trim, body, transmission), Laporan Project.



Kelompok: Car-a-thon

Stage: 2

Mentor: Stephanie

Pukul/ Tanggal: 21.00 WIB/ 20 Mei 2022

Poin pembahasan:

1. Data cleansing

- Cleansing dapat dibagi dalam beberapa stage, yaitu meninjau missing value, duplicated, outliers.
- Sebelum drop kolom, coba terlebih dahulu untuk reduksi unique value atau melihat hubungan dengan fitur utama. Jika drop fitur harus ada alasan yang tepat (Contoh: model, seller, trim).

2. Handle missing values

- Agar dapat mengakomodasi proses iteratif adalah dengan handling caranya. Disarankan untuk membuat bermacam-macam processing agar menghasilkan beberapa dataframe. Semua jenis dataframe dicoba untuk dimasukkan ke dalam model sehingga dapat mengetahui best practicenya.
- Jika masing-masing missing value tiap fitur jumlahnya sedikit, perlu dicek apakah secara keseluruhan fitur jumlahnya besar atau tidak. Jika data akhir yang didrop >20% lebih baik dicek ulang karena dapat mengurangi peluang dalam membuat model yang optimum.

3. Handle outliers

- Treatment dapat dilakukan dengan: 1) drop jika persentasenya tidak banyak; 2) dapat melakukan *cap* jika persentasenya cukup banyak; 3) dibiarkan saja.
- *Cap*: Jika outlier terdapat di atas pagar atas, maka perlu mengubah batas atas menjadi di pagar atas. Hal tersebut juga diterapkan pada outlier yang berada di bawah pagar bawah.

4. Feature transformation

- Log untuk mentransformasikan data yang skewed. Namun, fitur 'sellingprice' adalah target. Jika di log/transformation apapun, angkanya dapat berubah sehingga jika hendak membuat conclusion agak sulit karena angkanya berubah.
- Standardization adalah untuk membuat data menjadi distribusi normal. Tujuannya agar skalanya menjadi kecil (-3 sampai 3). Normalization (MinMaxScaler) adalah untuk mengubah skala namun valuenya antara 0 s/d 1 atau -1 s/d 1. Namun, MinMaxScaler sangat terpengaruh dengan outliers.

5. Feature encoding

- Jika sudah dibuat top 10 maka dianggap ada tingkatan sehingga dapat menggunakan label encoding.
- Rule of thumb jumlah kolom total adalah akar dari jumlah seluruh baris (setelah handling).

6. Feature engineering → sudah ok

7. Class imbalance

- Imbalance: data kategorikal dan setiap datanya imbalance (co: fraud → yang tidak fraud jauh lebih banyak dari yang tidak fraud).
- Hal yang perlu diperhatikan adalah classnya / target prediksi. Dalam penentuan harga mobil maka data merupakan continuous variables sehingga tidak akan imbalance.
- Evaluation metrics sudah banyak yang dapat menghandle class imbalance sehingga tidak selalu perlu dilakukan di awal (sebelum dimasukkan ke dalam model).



Kelompok: Car-a-thon

Stage: 2

Mentor: Stephanie

Pukul/ Tanggal: 21.00 WIB/ 20 Mei 2022

Hasil Diskusi:

- Apakah fitur dengan unique values yang banyak dapat langsung didrop?
Hati-hati dalam melakukan drop pada fitur yang memiliki unique values yang banyak, perlu dicek terlebih dahulu hubungannya dengan target fitur. Dapat dilihat dari hubungannya terhadap rentang 'sellingprice' (low, medium, high) atau mengurutkan dengan top 10.
- Kapan harus menggunakan modus atau median saat mengisi missing value?
Modus dapat digunakan untuk data kategorikal, sedangkan data numerik dan kontinu dapat menggunakan median atau mean (amannya menggunakan median, karena jika datanya skewed hasilnya tidak akan optimum jika menggunakan mean).
- Perbedaan handling outlier menggunakan beberapa metode?
Terdapat beberapa rumus yang mendefinisikan outlier: 1) Boxplot → asumsi data distribusi normal (melihat nilai z pada distribusi normal → mean=0, standar deviasi=1); 2) z score → perlu normalisasi dulu baru dilakukan z score (-3 sampai 3 adalah 99% data yang ada); 3) IQR → biasanya outliernya terlihat lebih banyak karena menggunakan batas pagar (datanya tidak jauh dari mean-nya), kelebihannya dapat menghandle data yang tidak normal.
- Apakah outlier perlu dihandle di awal atau nanti saat dimodelling?
Sebaiknya buat 2 jenis dataframe (1 outlier dibuang, 1 outlier tidak dibuang).
- Apakah untuk 'body' dan 'color' setelah diurutkan seperti 'make' boleh menggunakan label encoding (ketimbang one hot encoding)?
Bisa, selama encoding yang dilakukan memiliki arti. Untuk fitur 'make' jika belum diubah jadi top 10 namun sudah dibuat label, angka labelnya tidak dapat memperlihatkan urutan tersebut. Namun, jika sudah dibuat top 10 artinya label angka 0-9 sudah merepresentasikan urutan top 10 tersebut. Jika tidak berurutan, maka harus pake one hot encoding.

Tindak Lanjut:

- Buat beberapa opsi dataframe untuk dapat digunakan dalam model machine learningnya (mis: data yang dihandle missing value/outliersnya, data yang tidak dihandle missing value/outliersnya).
- Cek jumlah persentase total data yang didrop karena adanya missing value. Jika lebih dari 20%, maka harus ada yang dihandle bukan dengan cara drop.
- Cek kembali transformation yang dilakukan pada fitur target ('sellingprice'), transformasi yang dilakukan jangan sampai mengubah angka dan menyulitkan dalam membuat conclusion.
- Buat urutan terlebih dahulu untuk fitur-fitur yang akan diencoding sehingga alasannya jelas ketika menggunakan label encoding.
- Cek jumlah kolom akhir setelah melakukan semua proses. Jumlah kolom tidak boleh melebihi akar dari seluruh baris setelah pemrosesan.

Stage 3



Kelompok: Car-a-thon

Stage: 3

Mentor: Stephanie

Pukul/ Tanggal: 20.30 WIB / 27 Mei 2022

Pembagian tugas di stage ini:

Nama: Bagus Ganjar Lugina

Tugas: Linear Regression, LightGBM, Source Code, Laporan Project, Notulen Mentoring, Materi PPT.

Nama: Bernadetha Stella

Tugas: Decision Tree, XGBoost, Source Code, Laporan Project, Notulen Mentoring, Materi PPT.

Nama: Samuella Magdalena E

Tugas: SVR, LightGBM, Source Code, Laporan Project, Notulen Mentoring, Materi PPT.

Nama: Raihan Kurniasugianto

Tugas: Random Forest, Source Code, Laporan Project, Notulen Mentoring, Materi PPT.

Nama: M Rifqi Sarosa

Tugas: XGBoost, Source Code, Laporan Project, Notulen Mentoring, Materi PPT.



Kelompok: Car-a-thon

Stage: 3

Mentor: Stephanie

Pukul/ Tanggal: 20.30 WIB / 27 Mei 2022

Poin pembahasan:

1. Split data train-test

- Buat histogram train dan test agar dapat dicek apakah datanya benar-benar random (apakah bentuknya mirip antara data train dan test),
- Heatmap bukan untuk menentukan drop feature, spearman correlation pada heatmap adalah untuk mengukur korelasi secara linear. Sedangkan, feature bisa berkorelasi secara kuadrat, dll. Korelasi linear hanya sebagai proxy, bukan dibuang,
- Seberapa penting feature importance tergantung dengan modelnya (3 based model LGBM Selection, Boruta -> ngukur seberapa penting feature terhadap target).

1. Modelling

- Train yang jauh lebih bagus dari test artinya overfit dan memiliki variansi rendah, jika digunakan pada data lebih luas, sehingga modelnya akan menjadi kurang baik,
- Tentukan priority metrics, apakah akurasi atau yang lainnya. Sedangkan metrics lainnya menjadi pendukung dan harus ada alasannya,
- Akurasi regresi : 0.54 masih agak rendah. Paling mudah membaca MAE, jika $MAE=0.52$ artinya errornya cukup besar untuk data dengan rentang 0-1.
- Decision tree sudah dilakukan hyperparameter. Model tersebut terlihat overfitting (mencapai 30%) dan dapat dianggap lebih buruk dari linear regression.
- Untuk hyperparameter tuning pada random forest, terdapat 2 cara: 1) Randomizedsearchcv : put hyperparameter, ada beberapa kombinasi dan kemungkinan yang bisa terjadi, tidak dilakukan secara random semua dan ada maximum/end iterasinya (dapat ditulis misal 15x saja iterasinya) agar dapat stop running; 2) Gridsearchcv: mencoba semua iterasi yang mungkin.
- Untuk model random forest, RMSE dan MAE tidak overfit dan sudah cukup baik, errornya juga tidak terlalu parah.
- Untuk data yang cukup banyak model linear regression kurang baik, gradient boosting atau light GBM atau adaboost (untuk kategorikal) lebih baik digunakan untuk data yang cukup besar. Hasilnya juga cukup stabil dan dapat menghindari overfit, serta dapat menggunakan cukup banyak fitur. Infinite loop untuk hyperparameter tuning di XGBoost bisa terjadi karena scoringnya ga tepat sehingga bisa mengiterasi terus, dapat dicoba dengan mengganti scoringnya atau codenya.
- SVR biasa digunakan untuk classification, metodenya mirip KNearestNeighbor, hasilnya tidak akan jauh dari linear regression.

1. Feature selection

- Umumnya menggunakan feature selector, cari di packagenya (di XGBoost atau LightGBM ada), atau coba menggunakan boruta. Dapat juga menggunakan feature importance (recursive feature engineering) untuk melihat misal top 10, kemudian dicoba lagi menggunakan top 5, dst. Iterasi model berhenti ketika model tidak bertambah baik.



Kelompok: Car-a-thon

Stage: 3

Mentor: Stephanie

Pukul/ Tanggal: 20.30 WIB / 27 Mei 2022

Hasil Diskusi

- Bagaimana sebaiknya menentukan feature untuk model?

Best practice sebenarnya adalah menentukan final feature untuk dimasukkan ke model, namun jika belum yakin dapat melihat dulu feature importancenya.

- Hyperparameter tuning yang cocok untuk model didasarkan pada apa?

Hyperparameter tuning untuk membuat error semakin kecil. Model statistik memiliki hypermeter yang diset dari awal. Hyperparameter artinya parameter yang tidak berubah walau diberikan treatment apapun. Machine learning mencoba membaca data dan menentukan rumus awal. Tuning artinya mencoba melakukan loop atau iteration terhadap data tersebut. Nilai alpha dan solver dapat berubah-ubah bergantung pada data tersebut dianggap paling baik seperti apa. Linear regression hyperparameternya sangat sedikit. Untuk linear regression hyperparameter tidak terlalu banyak, namun untuk gradient boosting hyperparameternya dapat mencapai sekitar 20.

Tindak Lanjut:

- Menggunakan end iteration untuk model random forest dalam melakukan hyperparameter tuningnya.
- Mengubah scoring pada hyperparameter tuning untuk model XGBoost.
- Menggunakan model gradient boosting atau lightGBM sebagai perbandingan untuk mencari model yang lebih sesuai.
- Menggunakan feature selector untuk model XGBoost dan LGBM.
- Melakukan iterasi model sampai menemukan error yang semakin kecil dan jumlah fitur yang paling tepat untuk model.