Rakamin Academy

Homework EDA

Analytic Adventurers

Final Project

Estimasi Waktu Pengerjaan

3 - 5 jam

Jumlah Soal

5 Soal

Total Point

100 poin

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok, sesuai kelompok Final Project**

2. **Masing-masing anggota kelompok tetap perlu submit ke LMS** (jadi bukan perwakilan)

3. File yang perlu dikumpulkan:
   - File **jupyter notebook** (.ipynb) yang berisi source code. Tuliskan juga insights yang ditemukan ke notebook tersebut sebagai text markdown.
   - File **laporan homework** (.pdf) yang berisi rangkuman dari insight-insight yang diperoleh, beserta rekomendasinya (rekomendasi pre-processing untuk EDA, dan rekomendasi bisnis untuk business insight).

4. Upload hasil pengerjaanmu melalui LMS.
   - Masukkan semua file ke dalam **1 ffile** dengan format **ZIP**.
   - Nama File:

     `EDA - <Nama Kelompok>.zip`

1. Descriptive Statistics (15 poin)

Rakamin Academy

Gunakan function **info** dan **describe** pada dataset final project kalian. Tuliskan hasil observasinya, seperti:

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

Terdapat 4888 observasi (baris) dan 20 variabel (kolom) dalam dataframe. Dari 20 variabel tersebut: 15 di antaranya adalah kategorikal. 5 adalah numerik. Tidak ada kolom yang tergolong sebagai kategorikal yang seharusnya numerik (cat_but_car). Ada 9 kolom numerik yang memiliki jumlah nilai unik kurang dari batas ambang 10 dan seharusnya kategorikal (num_but_cat).

B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Terdapat 7 kolom/fitur yang memiliki nilai kosong, yaitu kolom Age (226), TyopeofContact (25), DurationOfPitch (251), NumberOfFollowups (45), PreferredPropertyStar (26), NumberOfTrips (140), NumberOfChildrenVisiting (66), dan MonthlyIncome (133).

C. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

Terdapat beberapa kolom/fitur yang memiliki data kosong (seperti yang sudah disebutkan pada poin B). Pada fitur DurationOfPitch, NumberOfFollowups, NumberOfTrips, dan MonthlyIncome memiliki nilai maksimum yang cukup tinggi, sehingga perlu diteliti apakah ada outlier atau tidak.

2. Univariate Analysis (25 poin)

Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus di-follow up saat data pre-processing.

- Kolom CustomerID memiliki sebaran data yang terlalu banyak sehingga kolom tersebut bisa dihapus nantinya.
- Kolom Age memiliki distribusi yang hampir normal.
- Kolom DurationOfPitch, NumberOfTrips, dan MonthlyIncome sepertinya memiliki distribusi data positive skewed yang mengindikasikan terdapat outlier, nantinya outlier dapat dihapus pada data training.
- kolom lain yang sisanya termasuk jenis data diskrit atau ordinal.

Selain itu, kesimpulan dari hasil analisis sebagai berikut:

- Customer dengan tipe kontrak Self Enquiry membeli paket lebih banyak daripada customer dengan tipe kontrak Company Invited
- Customer yang berada di city tier 3 memiliki persentase pembelian paket lebih tinggi setelah ditawarkan oleh sales
- Customer dengan Occupation Salaried dan Small Business memiliki ketertarikan untuk membeli paket yang ditawarkan
- Customer dengan gender Male lebih banyak mengambil paket yang ditawarkan daripada female atau fe male
- Distribusi jumlah orang yang ikut dalam perjalanan dengan customer yang mengambil penawaran paket travel paling banyak adalah 3 orang
- Customer yang di-follow up antara 3-5 kali lebih banyak yang mengambil penawaran travel dibandingkan dengan yang ditawarkan kurang dari 3 kali atau lebih dari 5 kali

2. Univariate Analysis (25 poin)

Selain itu, kesimpulan dari hasil analisis sebagai berikut (lanjutan):

- Product basic yang ditawarkan oleh sales lebih banyak diambil daripada produk lainnya
- Customer yang menerima penawaran paket travel lebih banyak memilih property bintang tiga dibanding bintang empat dan lima
- Customer dengan status single atau unmarried lebih banyak menerima penawaran paket travel
- Customer yang memiliki passport memiliki persentase menerima penawaran paket travel lebih tinggi daripada yang tidak memiliki paspor
- Customer yang memberikan score kepuasan >=3 lebih banyak membeli paket perjalanan.
- Customer yang memiliki mobil lebih banyak menerima penawaran paket travel
- Customer dengan jumlah anak 1 lebih banyak menerima penawaran paket travel
- Customer dengan jabatan Executive lebih banyak menerima penawaran paket travel

3. Multivariate Analysis (15 poin)

Lakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Tuliskan hasil observasinya, seperti:

A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

Fitur CityTier, DurationOfPitch, NumberOfPersonVisiting, NumberOfFollowups, PreferredPropertyStar, NumberOfTrips, Passport, PitchSatisfactionScore, NumberOfChildrenVisiting memiliki korelasi positif dengan ProdTaken. Fitur Age, OwnCar, dan MonthlyIncome memiliki korelasi negatif dengan ProdTaken.

Fitur yang berkorelasi kuat dengan ProdTaken pada bagian numeric terdapat DurationOfPitch, untuk numeric tapi kategori terdapat Passport dan untuk kategori terdapat ProductPitch serta Designation. fitur tersebut memiliki korelasi positif dan berkaitan kuat dengan fitur target.

B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?
kolom NumberofPersonVisiting dan kolom NumberofChildrenVisiting memiliki korelasi yang sangat tinggi (0.61) dan ada kemungkinan kolom tersebut redundant dan akan dipilih salah satu.

Pada feature kategori yang telah dilakukan one-hot encoding, banyak menghasilkan kolom redundan dan hal ini disebut **Dummy Variable Trap,** dikarenakan memasukan semua hasil one-hot encoding, untuk menghindarinya dapat di drop salah satu kolom hasil one-hot encoding

4. Business Insight (30 poin)

Selain EDA, lakukan juga beberapa analisis dan visualisasi untuk menemukan suatu business insight. Tuliskan minimal 3 insight, dan berdasarkan insight tersebut jelaskan rekomendasinya untuk bisnis.

1. **Product mana yang diminati oleh pelanggan?**

   Dari diagram batang dapat dilihat bahwa 60% paket yang diambil tawarannya adalah pake Basic, kemudian Deluxe, Standard, serta Super Deluxe dan King. Jika ditelaah lebih dalam pada setiap paketnya, tetap lebih banyak paket Basic yang diambil setelah ditawarkan, kemudian paket standar, paket deluxe, paket king, dan terakhir adalah paket super deluxe.

   Dalam presentasi ini, harga paket bukanlah masalah utama seseorang dalam mengambil paket setelah ditawarkan asalkan sales dapat dengan tepat memilih segmen pelanggan yang disesuaikan untuk di-pitch.

2. **Apakah Customer dari tiap City Tier yang berbeda memiliki ketertarikan dalam membeli paket perjalanan?**

   Adanya kecenderungan bahwa semakin banyak jumlah follow-up yang dilakukan oleh sales, semakin tinggi persentase pelanggan yang akhirnya memutuskan untuk membeli paket perjalanan. Persentase pelanggan yang membeli meningkat secara signifikan dari sekitar 11.36% pada satu follow-up menjadi 39.71% pada enam follow-up.

   Hal ini menunjukkan adanya korelasi positif antara jumlah follow-up yang dilakukan oleh sales dengan keputusan pelanggan untuk membeli paket perjalanan. Semakin banyak interaksi atau tindak lanjut yang dilakukan, semakin tinggi kemungkinan pelanggan untuk mengambil keputusan pembelian. Hal ini dapat menjadi indikasi bahwa strategi follow-up yang lebih intens memiliki dampak yang positif terhadap peningkatan konversi pelanggan dalam membeli paket perjalanan.

4. Business Insight (30 poin)

3. **Apakah Customers yang memiliki passport lebih tertarik mengambil paket perjalanan?**

   Tingkat konversi, atau persentase pelanggan yang telah mengambil produk, ternyata lebih tinggi di CityTier 3 (23.60%) dan CityTier 2 (23.23%) meskipun jumlah pelanggannya lebih sedikit dibandingkan dengan CityTier 1 (16.30%) yang memiliki jumlah pelanggan terbanyak. Hal ini menunjukkan bahwa meskipun jumlah pelanggan bisa lebih sedikit, tingkat konversi yang lebih tinggi di CityTier 3 dan 2 bisa menjadi peluang yang menarik dalam strategi pemasaran, mungkin dengan fokus lebih lanjut pada profil atau preferensi pelanggan di tingkat kota tersebut.

   terdapat perbedaan signifikan dalam persentase pelanggan yang telah mengambil produk antara mereka yang memiliki paspor dan yang tidak memiliki. Proporsi pelanggan yang telah mengambil produk jauh lebih tinggi di antara mereka yang memiliki paspor (34.74%) dibandingkan dengan yang tidak memiliki (12.29%). Hal ini bisa menunjukkan adanya korelasi atau pengaruh antara kepemilikan paspor dengan keputusan pelanggan untuk mengambil produk. Namun, perlu analisis lebih lanjut untuk memahami apakah faktor kepemilikan paspor secara langsung mempengaruhi keputusan tersebut atau terdapat faktor lain yang turut berperan dalam pengambilan keputusan pelanggan.

4. Business Insight (30 poin)

4. **Segmentasi Umur dan Income apa yang mendominasi di setiap paket perjalanan?**

Dilihat dari diagram batang, dapat diambil kesimpulan bahwa setiap paket memiliki segmen pasar (khususnya dalam income) sesuai dengan harga paket tersebut. Paket termurah yaitu Basic didominasi oleh pelanggan dengan low income. Paket menengah yaitu Deluxe didominasi dengan pelanggan dengan middle income. Paket menengah ke termahal yaitu paket Standard, Super Deluxe, dan King didominasi oleh pelanggan dengan high income.

Dikarenakan paket baru yang akan ditawarkan merupakan paket yang berfokus pada kesehatan yaitu paket Wellness Tourism dan dengan biaya/tarif yang lebih mahal dibandingkan dengan paket King, sehingga kedepannya pemasaran paket paket Wellness Tourism bisa didasarkan oleh segmentasi income dari King Package.

Dilihat dari diagram batang, paket Basic didominasi oleh anak muda dan orang tua, sedangkan paket Deluxe dan Super Deluxe didominasi oleh umur menengah, dan paket Standard dan Paket King didominasi oleh orang tua. Jika pemasaran paket Wellness Tourism mengacu pada paket King, maka segmentasi umur yang sesuai adalah orang tua berumur 50 tahun ke atas dengan high income.

5. Git (15 poin)

Link Github : https://github.com/reynaldi15/Rakamin-Analytic-Adventurers

Rakamin Academy

Selamat Mengerjakan!