Halo Semuanya, nama saya Evan, hari ini saya akan menjabarkan steps atau cara untuk membuat sebuah model machine learning random forest regressor, dari data analysis, visualisasi, training, evaluating, hingga prediksi dengan data baru.

Pertama saya akan memberikan sedikit overview untuk membuat teman-teman lebih paham tentang pembuatan model dalam video ini.

Dataset yang digunakan adalah dataset, Restaurant Sales Datasets

Restaurant Sales Datasets adalah dataset yang saya dapatkan dari CV Balibul untuk dianalisa dan digunakan untuk membuat model Prediksi

Model yang akan digunakan adalah model Random Forest Regressor dari library sklearn.

Apa sih Random Forest Regressor itu?
RANDOM FOREST REGRESSOR

Random Forest Regressor adalah algoritma Machine Learning yang termasuk dalam kategori Ensemble Learning dan digunakan untuk memecahkan masalah regresi. Ini adalah variasi dari algoritma Random Forest, yang pada dasarnya merupakan kumpulan dari beberapa pohon keputusan (Decision Trees). Random Forest Regressor dibangun dengan menggabungkan prediksi dari beberapa pohon keputusan untuk menghasilkan prediksi regresi yang lebih akurat dan stabil.

Sebelum kita masuk ke step yang lebih lanjut, kita akan melakukan import pada library-library yang akan kita gunakan.

Nah, pertama kita akan melakukan EDA terlebih dahulu, jadi apa sih EDA itu?

EDA adalah singkatan dari "Exploratory Data Analysis" atau Analisis Data Eksploratif dalam bahasa Indonesia. Ini adalah suatu pendekatan untuk menganalisis dataset dan merumuskan hipotesis berdasarkan pemahaman awal terhadap data. Tujuan utama dari EDA adalah untuk memahami karakteristik utama dari data, mengidentifikasi pola yang menarik, dan menentukan langkah-langkah berikutnya dalam analisis.

Nah, untuk EDA pertama, mari kita mencoba memasukkan datasets kita dari excel ke bentuk dataframe dalam python.

Setelah itu kita coba melakukan cek isi data yang kita punya.

Karena Data yang ada merupakan data tiap item dalam sebuah transaksi, kita akan mengumpulkan data ke tiap struk menggunakan pandas menggunakan SUM, berikut adalah cara yang saya lakukan.

Sekarang kita coba untuk melihat data kita sekarang, kita sudah mendapatkan data tiap struk yang ada, untuk melakukan analisis penjualan, kita dapat menggunakan data ini untuk mendapat data harian, mari kita coba melakukan sum Kembali untuk subtotal dan transaksi pada tiap shift di setiap harinya, selain itu kita juga akan menambahkan data hari untuk analisis perbedaan penghasilan setiap weekdays dan weekends.

Untuk menghilangkan hari tanpa transaksi, disini kita juga akan mencoba untuk menghilangkan subtotal dengan nilai 0, dan melakukan analisis describe untuk melihat data yang kita miliki,

Step berikutnya adalah Data Visualization, pada Visualisasi pertama, saya akan membuat sebuah histogram atau diagram batang untuk perbandingan antara shift 1 dan shift 2

Dari Visualisasi diatas, terlihat data subtotal shift 2 lebih tinggi dari shift 1, sehingga shift mempengaruhi Subtotal dan dapat dijadikan features.

Berikutnya, saya akan mencoba analisis pada hari menggunakan 2 plot, yaitu histogram dan swarm plot, untuk hal ini saya akan menggunakan chat gpt untuk membantu saya dalam pembuatan plot dengan ipywidgets

Dari Plot Histogram dan swarm plot, terlihat hari berpengaruh pada subtotal, pada weekend, terlihat penyebaran subtotal yang berada pada level tinggi (Pendapatan tinggi), sehingga hari dapat menjadi feature.

Step Berikutnya adalah Training

Dalam Training, kita perlu melakukan data Preprocessing terlebih dahulu, Data preprocessing adalah langkah kritis dalam proses machine learning untuk mempersiapkan data sebelum dilakukan training.

Pertama, kita akan melakukan importing untuk training dan evaluasi.

Setelah itu, Karena kita akan menggunakan RF Regressor, maka variabel kita harus angka, maka disini kita mengganti variabel hari kedalam bentuk angka

Setelah itu, kita akan menentukan features, features sudah kita dapatkan pada visualisasi dan EDA, dimana didapatkan bahwa shift dan hari berpengaruh pada subtotal, kita juga akan melakukan splitting data training dan testing,

Untuk Train pertama, disini saya mencoba n_estimators 100 dan random state 42, yang berarti model membuat 100 decision tree dengan random state 42

Kemudian kita akan mencoba melakukan evaluasi model dengan melakukan prediksi pada X_test, dan menghitung mse,r2, dan mae, hasil yang didapatkan terdapat MSE dan MAE cukup tinggi, namun ini dikarenakan data kita merupakan data jutaan sehingga aka nada error yang tinggi, namun ini masih dapat diterima, untuk data kita, kita dapat melihat pada R Squared, dimana didapat R Squared 0.51 yang cukup baik

Untuk sedikit memperbaiki error, kita dapat menambah jumlah decision tree, disini saya mencoba menambah 10 decision tree menjadi 110, lalu kita coba evaluasi lagi

Didapatkan terdapat penambahan akurasi. Kita akan sedikit melihat ke R-Squared disini untuk model kita

R-squared, juga dikenal sebagai coefficient of determination, memberikan gambaran seberapa baik model regresi sesuai dengan data yang diobservasi. R-squared berkisar antara 0 dan 1, di mana:

R-squared = 1: Model sepenuhnya menjelaskan variasi dalam data.

R-squared = 0: Model tidak menjelaskan variasi dalam data sama sekali.

Dalam konteks R-squared:

0.7-1.0: Umumnya dianggap baik. Model secara efektif menjelaskan variasi dalam data.

0.5-0.7: Dapat diterima. Model memiliki kemampuan yang baik dalam menjelaskan variasi dalam data.

0.3-0.5: Rendah. Model mungkin memiliki beberapa nilai prediktif, tetapi masih banyak variasi yang tidak dijelaskan.

0-0.3: Sangat rendah. Model tidak efektif dalam menjelaskan variasi dalam data.

Selain itu kita akan mencoba melihat histogram residual untuk melakukan evaluasi model

Histogram Residual adalah selisih antara nilai aktual dan prediksi.

Dalam konteks prediksi model, histogram residual yang mendekat ke 0 dan memiliki frekuensi yang tinggi di sekitar nilai 0 adalah tanda yang baik. Ini menunjukkan bahwa model Anda memiliki kinerja yang baik dalam memprediksi data dan memiliki sedikit kesalahan yang signifikan.

MODEL KITA sudah siap, sekarang kita akan mencoba prediksi data baru, disini saya membuat data baru untuk 30 hari kedepan dalam shift 1 dan shift 2

Nahh, kita sudah mendapat hasil prediksi kita, sekarang saya akan mencoba menambahkan kolom prediksi ke data baru dan melihat hasil-hasil prediksi.

Berikut adalah hasil prediksi model kita, dengan menggunakan RF Regressor kita dapat membuat model untuk prediksi transaksi pada restoran dengan hasil yang cukup baik, sekian dari saya, terimakasih.