Rakamin Academy

Laporan Project

Stage 3

Modelling

Modelling Selection Selection

Dalam modelling dataset kali ini kami melakukan benerapa eksperiment yaitu :

1. Logistic Regression
2. KNN
3. XG Boost
4. Adabosst
5. LightGBM

Modelling

LightGBM

Langkah awal adalah split Feature dan Label

```
[ ]  # Split Feature and Label
     X = df_clean[['hotel','adults','is_repeated_guest','previous_cancellations','previous_bookings_not_canceled','reserved_room_type','assigned_room_type','booking_changes','days_in_waiting_list',
      'required_car_parking_spaces','total_of_special_requests','lead_time_norm','adr_norm',
      'distribution_channel_Corporate',
      'distribution_channel_Direct',
      'distribution_channel_GDS',
      'distribution_channel_TA/TO',
      'distribution_channel_Undefined',
      'deposit_type_No Deposit',
      'deposit_type_Non Refund',
      'deposit_type_Refundable',
      'customer_type_Contract',
      'customer_type_Group',
      'customer_type_Transient',
      'customer_type_Transient-Party','total_stays',
      'total_guests','kids','arrival_date_year',
      'arrival_date_week_number_norm',
      'arrival_date_day_of_month_norm','guest_location','meal','market_segment',
     ]]

     y = df_clean['is_canceled'] # target / label

     #Splitting the data into Train and Test
     from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)
```

Modelling

LightGBM

Modeling menggunakan default parameter didapat score sebagai berikut:

- Accuracy (Test Set): 0.8641
- Precision (Test Set): 0.8494
- Recall (Test Set): 0.7716
- F1-Score (Test Set): 0.8086
- AUC: 0.9417

Dapat dilihat kita mendapatkan score presisi sebesar 84% yang menjadi acuan metrics

- Training set score: 0.8709
- Test set score: 0.8641

Terlihat jika skor training model LightGBM sebesar 87% sangat dekat dengan skor testing 86,4%, yang berarti bahwa model tidak overfit atau underfit .

Modelling

Logistic Regression

Modeling menggunakan default parameter didapat score sebagai berikut:

- Accuracy (Test Set): 0.80
- Precision (Test Set): 0.81
- Recall (Test Set): 0.62
- F1-Score (Test Set): 0.70
- AUC: 0.88

Dapat dilihat kita mendapatkan score presisi sebesar 81% yang menjadi acuan metrics

- Train score: 0.8062223285868134
- Test score:0.8031605986151441

Terlihat jika skor training model logistic regression sebesar 80.6% sangat dekat dengan skor testing 80.3%, yang berarti bahwa model tidak overfit atau underfit .

Model Evaluation

Algorithm	Precision (Test)	Accuracy (Test Set)	Recall (Test Set)	F1-Score
Logistic Regression	81%	80%	62%	70%
XGBoost	83%	84%	71%	77%
KNN	79%	84%	76%	78%
Adaboost	81%	82%	68%	74%
LightGBM	86%	88%	81%	83%

Sebagaimana dinyatakan dalam pernyataan masalah kami, kami mencari model yang memberikan nilai presisi tertinggi sebagai acuan metrics. Model terbaik dari hasil evaluasi yang dipilih adalah LightGBM yang memberikan performa terbaik.

Hyperparameter Tuning

LightGBM

Kami mencoba hyperparameter tuning dengan parameter *objective, metric, num_boost_round, dan learning rate*. Nantinya optuna akan mencoba coba kombinasi parameter yang tepat dan melakukan trial sejumlah yang kita masukkan (disini n_trial yang digunakan 100).

Dengan menggunakan hyperparameter tuning didapat score sebagai berikut:

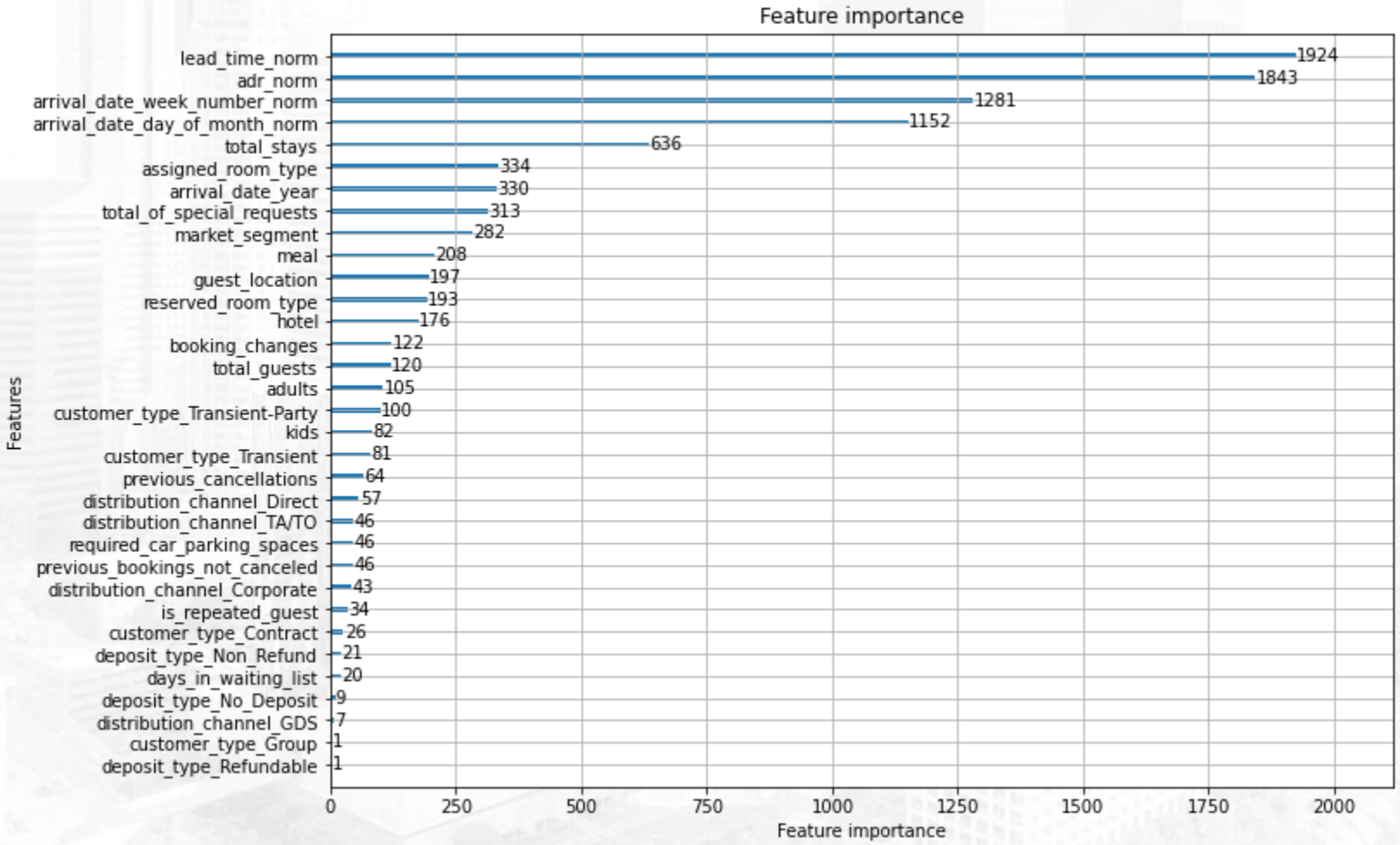
- `Accuracy (Test Set): 0.8812`
- `Precision (Test Set): 0.8607`
- `Recall (Test Set): 0.8123`
- `F1-Score (Test Set): 0.8358`

Dapat dilihat setelah melakukan hypertuning score presisi meningkat menjadi 86% dari nilai sebelumnya 84%

```
Training set score: 0.9335
Test set score: 0.8812
```

Terlihat jika skor training model LightGBM setelah hyperparameter tuning sebesar 93.3% dekat dengan skor testing 88,1%, yang berarti bahwa model tidak overfit atau underfit .

Feature Importance



Feature Importance

Top Feature

1. Lead Time
2. Adr
3. Arrival date week number
4. Arrival date day of month
5. Total stays

Business Insight & Recommendation

Dari hasil feature importance dapat terlihat bahwa Lead time mempunyai pengaruh terbesar dalam pembatalan pesanan hotel. Maka dari itu pihak hotel dapat menerapkan batasan maksimal waktu pemesanan kamar dan menerapkan doposit/uang muka untuk reservasi dengan jangka waktu lama guna menurunkan tingkat pembatalan pesanan.

Feature selection

Top Feature

Melakukan iterasi modeling dengan 10 top feature yang dipilih didapat score sebagai berikut:

- `Precision (Test Set): 0.8312`
- `Recall (Test Set): 0.7628`
- `F1-Score (Test Set): 0.7955`

Dapat dilihat setelah melakukan feature selection score presisi sebesar 83%

```
Training set score: 0.8852
Test set score: 0.8311
```

Terlihat jika skor training model LightGBM dengan feature selection sebesar 88.5% dekat dengan skor testing 83,1%, yang berarti bahwa model tidak overfit atau underfit .