Capstone Modul III

Oleh : Ian Arif Rahman

HOUSE PRICE PREDICTIONS IN CALIFORNIA

CONTENTS

Background

- Bertambah populasi
- Pembangunan perumahan bersifat ireversible
- Harga yang tidak cocok berpotensi tidak laku/ low margin
- Penetapan harga yang sesuai sesuai adalah kunci

Audience --> C- Level Developer

Problem --> menentukan harga rumah yang tepat dan sesuai

Goals --> mendapatkan prediksi harga rumah terbaik

Analytical Approach --> model regresi

Metric Evaluation -->Root Mean Squared Error (RMSE) & R-squared (R2)

Dataset

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	NEAR BAY
6	-122.25	37.84	52.0	2535.0	489.0	1094.0	514.0	3.6591	299200.0	NEAR BAY
7	-122.25	37.84	52.0	3104.0	687.0	1157.0	647.0	3.1200	241400.0	NEAR BAY
8	-122.26	37.84	42.0	2555.0	665.0	1206.0	595.0	2.0804	226700.0	NEAR BAY
9	-122.25	37.84	52.0	3549.0	707.0	1551.0	714.0	3.6912	261100.0	NEAR BAY

- Missing value data 207 dikolom total_bedrooms dan dari histogram terkonfirmasi bahwa distribusi data adalah positive Skewness, jadi kita putuskan untuk mengisi dengan nilai median.
- Kita memutuskan mempertahankan outliers kecuali yang sangat extreme seperti yang terjadi di housing_median_age & median_house_value

Preprocessing

Add Columns :

- Ruangan per rumah tangga (rooms_per_household)
- Ruang tidur per Ruangan (bedrooms_per_room)

tindakan yang akan dilakukan :

- longitude, latitude -> None
- housing median age, total rooms, total bedrooms, population, households, median income, rooms per household, bedrooms per room -> log transform
- ocean proximity -> label encoding

Scaling -> Standar Scaler

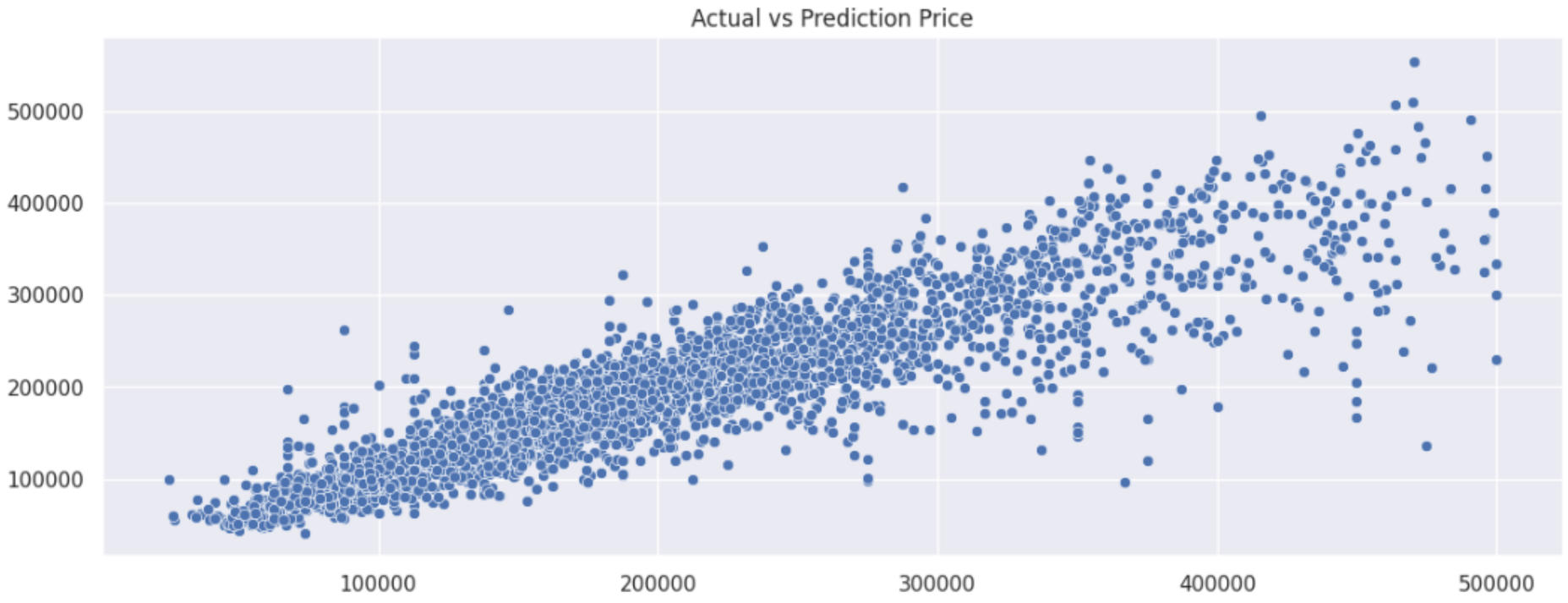
Data Modeling

	0	1	2	3	4	5	6
Model	Linear Regression	KNN Regressor	Random Forest	Catboost	XGBoost	LightGBM	Gradient Boosting
RMSE	0.313278	0.276569	0.215244	0.204017	0.20921	0.214466	0.251982
R-square	0.656412	0.276569	0.837803	0.854283	0.846769	0.838974	0.777711

Dari tabel diatas kita bisa melihat nilai RMSE dan R-square semua algoritma. Model terbaik menurut RMSE adalah yang bernilai paling kecil. RMSE terkecil dimiliki oleh Catboost senilai 0.204017. Sedangkan untuk R-square, model terbaik adalah ketika nilai R-squarenya paling besar yang juga sama yaitu Catboost dengan nilai 0.854283. Sehingga, kita bisa simpulkan model terbaik adalah Catboost

Actual vs Prediction Price

	Actual	Predicted
12869	133000.0	123378.19
8961	332800.0	305008.51
20309	194400.0	205956.15
17392	117600.0	135862.51
6961	198600.0	221112.45
...	...	...
6255	155800.0	150739.21
3759	174200.0	204438.56
859	247600.0	243243.48
18315	406300.0	340433.64



Conclusion

Dari proses yang sudah kita lakukan, kita bisa simpulkan untuk case dataset ini, 3 algoritma boosting relatif lebih perform dibandingkan dengan yang lain dan untuk random forest performanya cukup bisa bersaing.

Recomendation

- Bisa ditambahkan informasi lain yang berhubungan langsung dengan harga rumah seperti fasilitas, luas, developer rangking, dll
- Jika dimungkinkan, bisa update data terbaru. karena dataset yang diolah adalah data lama ditahun 1990. Tentu bisa jadi sudah ada bangunan baru disekitar lokasi dan membuat harganyapun jadi sudah tidak relevan dengan keadaan sekarang.
- Untuk model, bisa dilakukan tuning agar hasilnya lebih baik lagi. Banyak pilihan yang tersedia yang bisa digunakan di hyperparameter tuning.
- Memprediksi harga rumah tentu sangat terpengaruh dengan waktu karena harga rumah relatif naik setiap tahun, sehingga perlu ada data yang berkesinambungan sedangkan ini ada gap sekitar hampir 30 tahun

Thank You