CODING DOJO

الأكاديمية السعودية الرقمية
SAUDI DIGITAL ACADEMY

VISION رؤية 2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

Capstone Project

PREDICTING THE NUMBER OF BUILDING PERMITS IN SAUDI ARABIA

content

Introduction

2030 Vision

a vibrant society, a thriving economy, an ambitious nation

Introduction

Problem statement and approach

Problem statement

I- Investors do not have a clear insight about where are the right locations for investing in real estate

Approach

- Predict the number of building permits in all regions of Saudi Arabia.
- Identify the best location for investment

content

Data Review

Data Source

Building Permits Dataset which was used in this case study has been taken from KAPSARC (King Abdullah Petroleum Studies and Research Center).

Variables and Explanations

Feature	Definition
Year	Date in years
Region	Region
Building type	Building type
Number of permits	Number of building permits
Total area of building	Total area of building in sqm
Total number of floors	Total number of floors
Total area of plot	Total area of plot in sqm
Total Length of fences	Total length of fences in lm
Longitude	Longitude
Latitude	Latitude

Data preprocessing

Data cleaning

1 - Missing Data
2 - Noisy Data
- Rename columns
- Outlier analysis
- Detect Multicollinearity with VIF

Data transformation

1- Feature engineering:
Label encoding

2- Feature scaling:
Standardized scaler

content

INTRODUCTION

2030 Vision
Problem statment

Data Review

Explaining the Dataset &
preprocessing techniques

Data Exploration

Tableau visualization

Results

Data modelling and the
results

Conclusion

Conclusion and Future
work

content

INTRODUCTION	Data Review	Data Exploration	Results	Conclusion
2030 Vision Problem statment	Explaining the Dataset & preprocessing techniques	Tableau visualization	Data modelling and the results	Conclusion and Future work

Choose the model

Target

Our target is to predict the **number of permits**, which means it is a regression problem.

Model

- Applied 15 deferent regression models and for measuring the models performance we used MAE (Mean Absolute error).
- Scaled the data with a standardized scaler and compared it with the default model.

Choose the model

- Chose the best 2 (with the lowest MAE)
- Evaluate the models with the Val set
- Hyper-parameter tuning for the 2 models

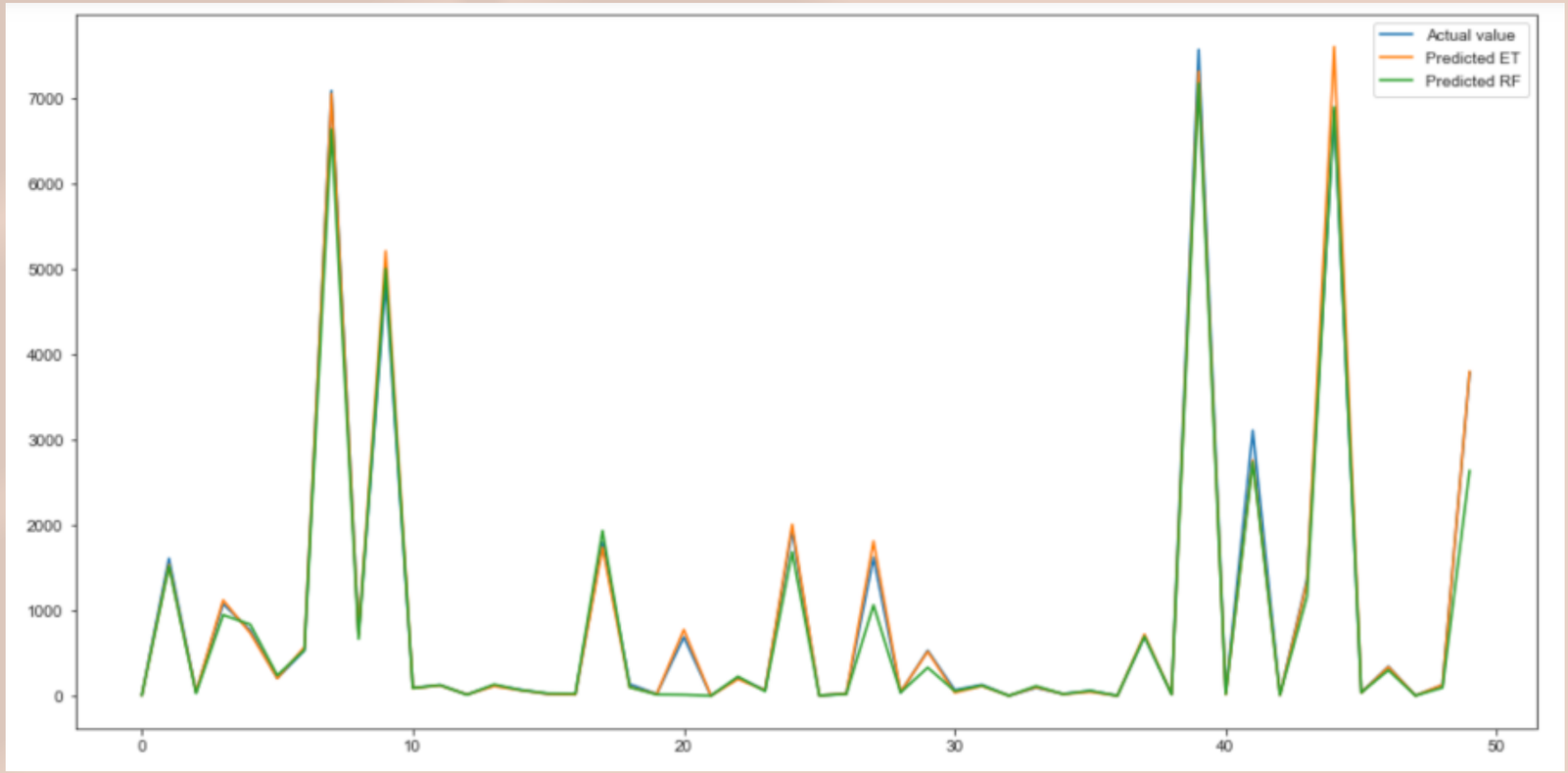
DEFAULT MODEL

	model	run_time	MSE	MAE	RMSE
5	ExtraTree	0.0	8.487632e+04	82.33959	291.33540
0	RandomForestRegressor	0.01	1.499975e+05	96.95865	387.29509
11	HuberRegressor	0.0	1.825589e+05	118.21323	427.26912
4	Bagging	0.0	1.772617e+05	118.41049	421.02459
1	DecisionTreeRegressor	0.0	1.912585e+05	137.19850	437.33115
14	Lasso	0.0	1.641140e+05	168.08248	405.10983
13	Ridge	0.0	1.641008e+05	168.21528	405.09357
10	LinearRegression	0.0	1.641008e+05	168.21718	405.09355
12	RANSAC	0.0	4.288625e+05	196.64019	654.87596
3	KNeighbors	0.0	3.803388e+05	274.20375	616.71612
6	AdaBoost	0.0	1.875317e+05	357.73883	433.04929
9	LinearSVR	0.0	8.786739e+05	383.25969	937.37606
7	SVR	0.0	2.763020e+06	664.60945	1662.23349
8	NuSVR	0.0	2.677040e+06	688.46174	1636.16615
2	GaussianProcessRegressor	0.0	3.058088e+06	738.35581	1748.73904

DEFAULT MODEL WITH SCALER

	model	run_time	MSE	MAE	RMSE
14	Lasso	0.0	3.292666e+05	573.81638	573.81756
2	GaussianProcessRegressor	0.0	1.697924e+06	624.43190	1303.04422
8	NuSVR	0.0	2.331990e+06	663.16835	1527.08530
7	SVR	0.0	2.409637e+06	682.05188	1552.30066
13	Ridge	0.0	2.423625e+06	692.89263	1556.79961
10	LinearRegression	0.0	2.425492e+06	692.93396	1557.39898
9	LinearSVR	0.0	2.821568e+06	695.53324	1679.75224
11	HuberRegressor	0.0	2.794004e+06	695.97372	1671.52737
1	DecisionTreeRegressor	0.0	2.864023e+06	712.88303	1692.34241
3	KNeighbors	0.0	2.983212e+06	718.43285	1727.19778
5	ExtraTree	0.0	2.887755e+06	727.61453	1699.33950
0	RandomForestRegressor	0.01	2.935894e+06	728.92831	1713.44516
4	Bagging	0.0	2.986128e+06	736.52985	1728.04180
12	RANSAC	0.0	4.420027e+06	835.08024	2102.38604
6	AdaBoost	0.0	3.025150e+06	957.57566	1739.29574

Evaluate the models



Hyper-parameter Tuning (Grid Search)

Extra Tree

Default	82.33959
Grid Search 1	117.54174
Grid Search 2	84.97646
Grid Search 3	106.196

Random Forest

Default	96.9856
Grid Search 1	150.10674
Grid Search 2	149.90262
Grid Search 3	144.39958

Results

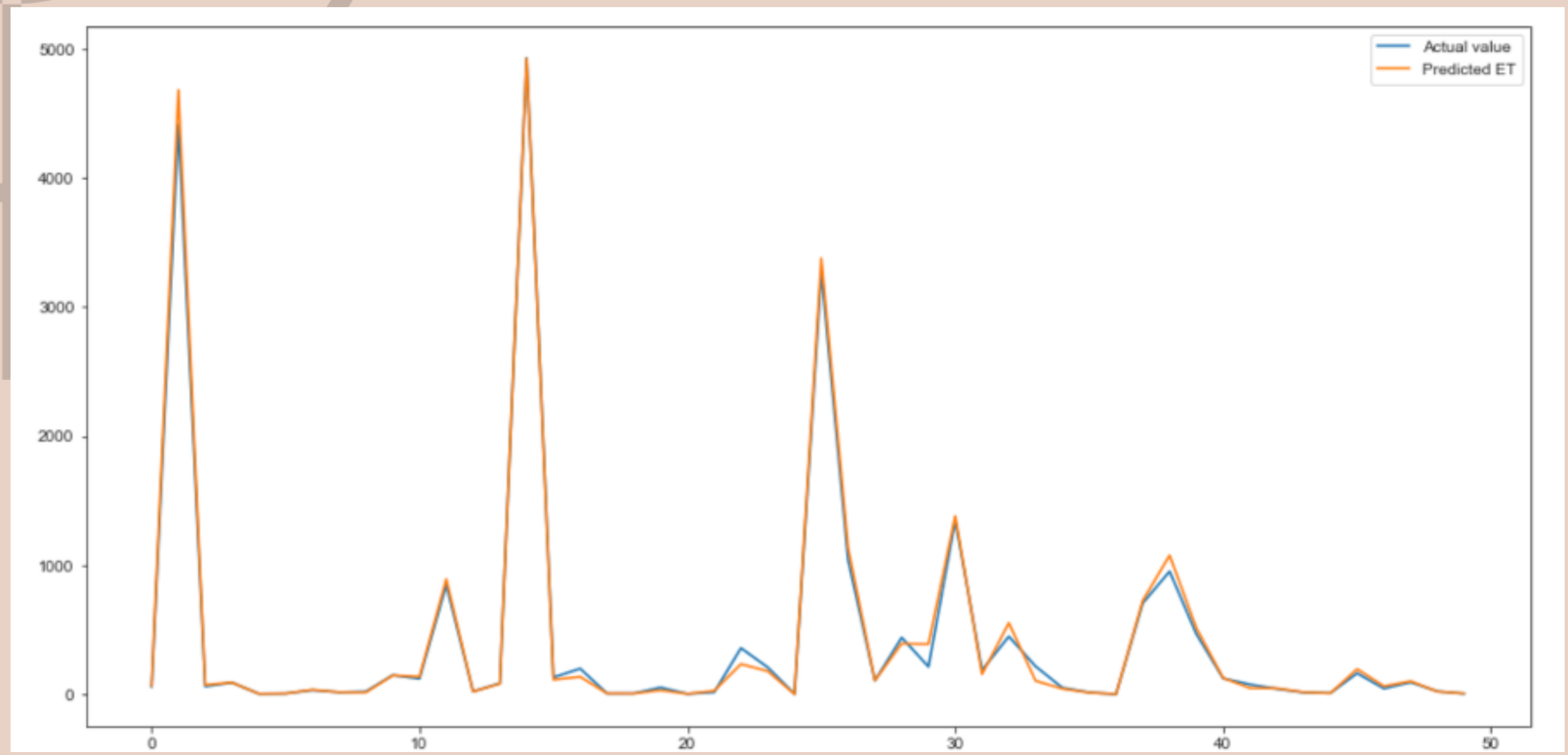
Prediction

- The prediction results on the test set

Error Precentage

- Calculate the error precentage

The prediction results



The result of the prediction using the Extra tree model, MAE: 101.64287

Error %

After making the prediction we calculated the error percentage, to find the number of points that have prediction error greater than 30%.

The number of records in the test set = 343

The number of points with error greater than 30% = 61

content

INTRODUCTION

2030 Vision
Problem statment

Data Review

Explaining the Dataset & preprocessing techniques

Data Exploration

Tableau visualization

Results

Data modelling and the results

Conclusion

Conclusion and Future work

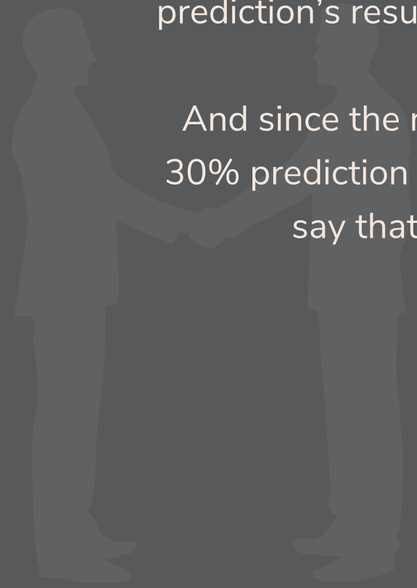
CONCLUSION

1- Applied 15 regression models to find the one with the best results

2- Scaled the data to see the effects of the scaler on the models

3- Calculated the error percent to see if our prediction's results are acceptable or not

And since the number of records with more than 30% prediction error is less than 20% then we can say that the results are acceptable.



FUTURE WORK

1- Extending our analysis by finding the optimum hyper-parameters for our machine learning models.

2- Applying feature selection approaches to determine the features most relevant to the number of permits.

3- Including supporting datasets in our analysis to help us further investigate the real estate market to provide better recommendations to the investors in this sector.

Thank you for listining

Any Question

By: Maram Al Shehri, Nouf Al Rehaili, Norah Al Harthi, Nourah Alsaadan