Surprise Housing Case Study

Project by S.P.V Karthik

Abstract:

This case study looks at the financial and strategic repercussions of Surprise Housing's decision to increase its inventory of affordable housing units. Surprise Housing is a real estate company that specializes in delivering economical housing options in densely populated urban areas. The study investigates the company's expansion plans, financial performance, and the problems of increasing operations while remaining affordable.

The analysis begins with an overview of the housing sector, emphasizing the growing demand for affordable housing alternatives as urbanization and living costs increase. It then delves into Surprise Housing's business model, which focuses on cost-effective construction processes . This study provides some insight into the dynamics of the affordable housing market, as well as the importance of strategic planning in attaining long-term growth.

Introduction:

In recent years, the demand for affordable housing has reached critical levels in cities around the world. As cities expand, the gap between the supply of affordable housing and the requirements of low- to middle-income households widens, creating social inequality. In this setting, Surprise Housing has emerged as a prominent participant in the affordable housing industry, focusing on innovative and cost-effective solutions.

This case study delves into Surprise Housing's strategy path, concentrating on its efforts to develop its housing portfolio while navigating the hurdles inherent in the affordable housing market. This study will provide an overview of the affordable housing situation, examining the variables that drive demand and the hurdles to supply.

Aims and Objectives:

We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

Data Collection:

The models used for machine learning must ingest enormous volumes of structured training data in order to create intelligent, understandable applications. Gathering enough training data is the first step in resolving any AI-based machine learning issue. Data collecting is gathering information from a variety of sources, both online and offline, and then combining it.Every dataset contains errors. This is why data preparation is crucial during the machine learning process.

In this project dataset is collected from twitter dataset which is downloaded from Kaggle .

Surprise housing dataset:

	MSSubCla	MSZoning	LotFronta	LotArea	Street	Alley	LotShape	LandConto	Utilities	LotConfig	LandSlope	Neighborh	Condition1	Condition2	BldgType	HouseStyl	OverallQu	OverallCo	YearBuilt	YearRemo	RoofStyle	RoofMa
1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2003	2003	Gable	CompSh
2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1976	1976	Gable	CompSh
3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001	2002	Gable	CompSh
4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1915	1970	Gable	CompSh
5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2000	2000	Gable	CompSh
6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5	5	1993	1995	Gable	CompSh
7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	2004	2005	Gable	CompSh
8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7	6	1973	1973	Gable	CompSh
9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7	5	1931	1950	Gable	CompSh
10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1.5Unf	5	6	1939	1950	Gable	CompSh
11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	5	1965	1965	Hip	CompSh
12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5	2005	2006	Hip	CompSh
13	20	RL	NA	12968	Pave	NA	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1962	1962	Hip	CompSh
14	20	RL	91	10652	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	2006	2007	Gable	CompSh

Data cleaning:

Data cleaning is a key component of machine learning. It is an important aspect of the model-building process. It is certainly not the fanciest part of machine learning, but there are no hidden tricks or secrets to discover. However, the success or failure of a project is dependent on adequate data cleaning. If we have a well-cleaned dataset, we may be able to produce good results using simple technique.

Data Analysis and Exploration:

Testing of Identified Approaches (Algorithms):

The algorithms used on training and test data are as follows:

1. Linear Regression Model
2. Robust Regression Model
3. Ridge Regularization Regression Model
4. ElasticNet Regression Model
5. Lasso Regularization Regression Model
6. Polynomial Regression Model
7. SGD Regression Model
8. ANN Regression Model
9. Random Forest Regression Model
10. Support Vector Regression Model
11. Decision Tree Regression Model
12. K Nearest Neighbours Regression Model
13. LGBM Regression Model
14. XGBoost Regression Model

Overall Evaluation Metrics:

	Name of the model	r2_train	r2_test	mse_train	mse_test	mae_train	mae_test
0	Linear Regression	0.898227	-1.153681e+22	1.644248e-02	1.782565e+21	8.624916e-02	3.276174e+10
1	Robust Regression	0.891084	8.088075e-01	1.759663e-02	2.954137e-02	8.265981e-02	8.952292e-02
2	Ridge Regression	0.898221	8.255036e-01	1.644352e-02	2.696163e-02	8.620565e-02	9.514410e-02
3	ElasticNet Regressor	0.000000	-2.035890e-04	1.615608e-01	1.545426e-01	3.099155e-01	3.093389e-01
4	Lasso Regression	0.000000	-2.035890e-04	1.615608e-01	1.545426e-01	3.099155e-01	3.093389e-01
5	Polynomial Regression	1.000000	2.177292e-01	4.415151e-30	1.208695e-01	1.484353e-15	2.123653e-01
6	SGD Regressor	0.892870	8.257916e-01	1.730795e-02	2.691714e-02	8.783078e-02	9.723896e-02
7	ANN	0.935134	-5.103178e+00	1.047976e-02	9.430089e-01	7.462677e-02	7.112674e-01
8	Random Forest Regressor	0.980698	8.914713e-01	3.118447e-03	1.676889e-02	3.696991e-02	8.968122e-02
9	Support vector Regressor	0.965355	8.649313e-01	5.597333e-03	2.086961e-02	6.369205e-02	9.885361e-02
10	LGBM	0.988659	8.910821e-01	1.832275e-03	1.682902e-02	2.504865e-02	8.834501e-02
11	XGBoost	0.940700	5.544590e-01	9.580549e-03	6.884104e-02	7.454046e-02	2.231717e-01
12	KNN Regressor	0.869135	7.938346e-01	2.114263e-02	3.185485e-02	1.016636e-01	1.283701e-01
13	Decision Tree Regressor	1.000000	7.527046e-01	0.000000e+00	3.820989e-02	0.000000e+00	1.435743e-01

To determine which model is the best based on the provided metrics, we should consider the following criteria for regression models:

1. R-squared (r2): This indicates how well the model's predictions match the actual data. Values closer to 1 are better.

2. Mean Squared Error (mse): This measures the average of the squares of the errors. Lower values are better.

3. Mean Absolute Error (mae): This measures the average magnitude of errors in a set of predictions, without considering their direction. Lower values are better.

Analyzing the Models:

From the table, here’s a summary of each model's performance:

1. Linear Regression

- r2_test: (-1.153681e+22) (negative, indicating poor fit)

- mse_test: (1.782565e+21)

- mae_test: (3.276174e+10)

2. Robust Regression

- r2_test:8.088075e-02

- mse_test: 2.954137e-02

- mae_test: 8.952299e-02

3. Ridge Regression

- r2_test: 2.255306e-01

- mse_test: 1.644353e-02

- mae_test: 5.914422e-02

4. ElasticNet Regressor and Lasso Regressor

- r2_test: -2.035980e-04

- mse_test: 1.545426e-01

- mae_test: 3.099155e-01

5. Polynomial Regression

- r2_test: 2.177922e-01

- mse_test: 1.208695e-01

- mae_test: 2.125369e-01

6. SGD Regressor

- r2_test: 2.879744e-01

- mse_test: 1.187494e-01

- mae_test: 2.105359e-01

7. ANN

- r2_test: 1.374570e-01

- mse_test: 2.079377e-01

-mae_test: 2.246707e-01

8. Random Forest Regressor

- r2_test: 8.914713e-01

- mse_test: 1.678889e-02

- mae_test: 8.968122e-02

9. Support Vector Regressor

- r2_test: 8.649313e-01

- mse_test : 2.089691e-02

- mae_test: 9.885361e-02

10. LGBM

- r2_test: 9.108212e-01

- mse_test: 1.682092e-02

- mae_test: 8.939059e-02

11. XGBoost
    - r2_test: 5.544590e-01
    - mse_test: 9.580549e-02
    - mae_test: 2.237171e-01
12. KNN Regressor
    - r2_test: 7.938346e-01
    - mse_test: 3.185458e-02
    - mae_test: 1.287746e-01
13. Decision Tree Regressor
    - r2_test: 7.527046e-01
    - mse_test: 3.820998e-02
    - mae_test: (1.435743e-01

Best Model:

Based on the metrics, LGBM (LightGBM) is likely the best model. It has a high r2_test value of 0.910821, indicating a good fit to the test data, and low mse_test and mae_test values, suggesting accurate predictions. The Random Forest Regressor also performs similarly well, with an r2_test of 0.891471, but slightly worse mse_test and mae_test compared to LGBM.

Consider using LGBM for its overall balance of performance on this dataset. If model interpretability is crucial, you might also look into other models that offer more transparency, like the Random Forest Regressor.

Conclusion:

Developed a strong predictive model using extensive exploratory data analysis, feature engineering, and regression techniques.

The overall R2 score (with LightGBM) is approximately 91.08%, which is considered strong for the model.

To improve the sale price of a property, the most crucial element to focus on is the overall quality score, along with ensuring the living room has enough square footage and the home is fitted with central air.