SURPRISE HOUSING CASE STUDY

SAKETH KUMAR NOOTHI

AGENDA:

- ABSTRACT
- OBJECTIVE
- INTRODUCTION
- METHODOLOGY
- CODE
- CONCLUSION

ABSTRACT

- We have a dataset about houses which has a shape of (1460,81) 81 columns that's huge! We now need to understand the variation of Saleprice that is Price of a house in Australia (houses in Australia) with different features we have all different kinds of data : continuous , discrete, categorical .now our company " Surprise Housing " has decided to rent or buy houses and start the business in Australia thereby increasing our revenue and diversification.
- We are now need to create a regression model for predicting sales price using Ridge And Lasso Regression find optimal regularization parameters. We use GridSerachCV CV stands for Cross Validation . Additionally, we need to determine the optimal values of alpha for Ridge and Lasso regression.
- We should build a regression model using regularization in order to predict the actual value of the prospective properties and decide to invest in them or not. The company wants to know the following things about the prospective properties: Which variables are significant in predicting the price of a house, and How well those variables describe the price of a house.
- This model will 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. Further, the model will be a good way for management to understand the pricing dynamics of a new market

OBJECTIVE:

- The primary objectives of this analysis and predictive modelling are :
  - Build regression models (Ridge and Lasso) to predict house prices based on available independent variables. Identify the significant predictor variables that influence house prices. Evaluate the models' performance and determine the optimal values of alpha for regularization. Assess the impact of optimal value of alpha value on the models and identify the most important predictor variables that contribute to Saleprice
  - We perform Date Preprocessing and Data Wrangling
  - For continuous , categorical variables we create a different data-frames in which categorical column are then converted to its dummy values i.e. one hot encoding
  - Then for continuous variables we scale them by using standardscale() object in python this is nothing but converting the continuous data in to normal distribution which also called Z-score normalization
  - There after we perform EDA exploratory data analysis where we analyze the correlation of numerical features with target variable and also perform multi variate analysis on continuous data to get the strong correlated columns or features
  - Then the we if the null values in Categorical Variables have any trends in saleprice thereby analyzing the variation in the distribution of variables in a particular column or feature.
  - After cleaning our data we train our model then plot the training error and test error
  - Thereby optimizing the models

INTRODUCTION:

- A US-based housing company, Surprise Housing, aims to expand its operations into the Australian market. To make informed investment decisions, the company collected a dataset containing various features related to house sales in Australia. This analysis aims to help Surprise Housing understand the factors that affect house prices and build predictive models for pricing. Key steps which includes data preprocessing, exploratory data analysis, feature engineering, feature selection and model building using ridge and lasso regression thereby producing the optimal value for hyperparameter lambda .

METHODOLOGY:

- Data Collection: The dataset containing information about house sales in Australia is taken of shape (1460,81).
- Data Preprocessing: Missing values are handled, data types are converted, and outliers are capped to its 5^{th} Percentile and 95^{th} percentile .
- Exploratory Data Analysis (EDA): Visualizations and statistical analysis are performed to understand the data characteristics and relationships between categorical and numerical features against saleprice .
- Data Exploration Univariate Analysis Bivariate Analysis , Data Treatment Dummy Variable Creation Feature Engineering: New features, such as ‘house age’ is created to enhance the modeling process and removing id column
- Data Scaling: Numerical features are standardized using StandardScaler (z score normalization).
- Model Building: Ridge and lasso regression models are constructed, and hyperparameters are tuned using cross-validation using ‘gridsearchcv’, Tuning & Evaluation Split the Data into Dependent and Independent variables Train - Test Split

- Model Evaluation: The models' performance is assessed using metrics such as R-squared, RMSE, and MAE .
- Coefficient Interpretation: Interpretation of coefficients is carried out to understand the impact of features on house prices.
- Calculating the betas values then find the best and highest values of betas to judge on the strong features that effect saleprice
- Inferences for 'Surprise Housing’ : Compare the two models outcomes in selecting the best features

CODE:

- https://drive.google.com/file/d/1zFgfPaYYub_guAUoZ6Eg_J5jjayzBUVo/view?usp=sharing

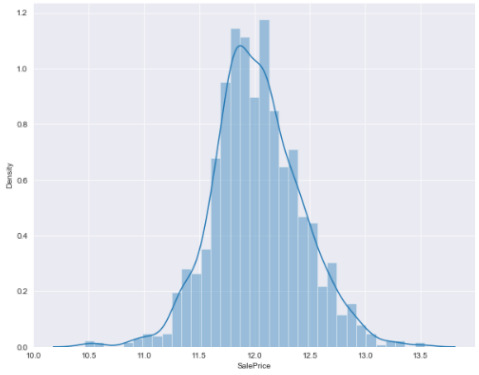
EXPLANATION: EDA AND PREPROCESSING

- Firstly after seeing the data I realized there 81 column in which there are only a few columns with missing values and many with NA values in it
- Then i replaced na values with none values to do analysis because in categorical variables the columns with null values indicate that there no such feature for ex in alley there na values mean that house doesn't have a alley near it
- Some of feature have wrong data type Ex mssubclass Identifies the type of dwelling involved in the sale it was assigned as numerical we convert it object
- And for numerical columns we convert them object to numeric ex lotfrantage we convert it to numeric using pd.to_numeric(df['LotFrontage'], errors='coerce') errors = coerce means that while converting the values to numeric the pandas treats a missing as nan values then converts to NAN values (not an number)
- Then we perform univariate analysis on numerical column we observe that Saleprice is skewed now we use log transformation to convert it to normal distribution

-> Skewed

```
In [97]: df['SalePrice'] = np.log(df['SalePrice'])

In [98]: ## First Let's analyse the target variable 'SalePrice'
         plt.figure(figsize=[10,8])
         sns.set_style('darkgrid')
         sns.distplot(df['SalePrice']);
```



->unSkewed

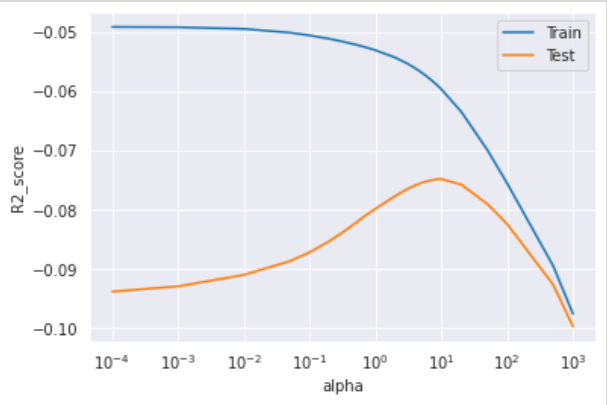
- We need to convert the target variable to normal distributed just to get rid of outliers while modeling ML models genrally perofrms best on un- skewed data or normally distributed data as input and output
- Next step is perform Univariate analysis on continuous data plotting the distribution in histogram and Boxplot for better understanding the skewness and outliers in the features.
- The above step helps us which all features are strongly correlated to saleprice and found that many features actually contains outliers and skewed right or left
- And target variable SalePrice is highly correlated with **GrLivArea, GarageCars and GarageArea**.
- For categorical I plotted the distribution plot then found out from value counts and bar plots that there is no column with single unique value that we could remove.
- After plotting the saleprice with categorical variable that is bivariate analysis found some column like 'Overall qual' had a trend we straight away say that this important feature but do not know which specific unique value of feature is important we do find this after using lasso and ridge regression
- After this multivariate analysis in numerical columns found out that Many features are highly correlated with each other. Target SalePrice is highly correlated **with GrLivArea, GarageCars and GarageArea.** As mentioned in eda process also !!!
- We can now reduce the shape by doing feature engineering we subtract yrsold , yrbuild to get age of the house
- After converting to numeric column we observed there missing value in lotfrontage , masvnrArea we can impute it with mean
- Now we can create dummy variables for categorical features which commonly known as one hot encoding which will increase the shape from (1460, 47) , (1460, 353) 47 to 353 so 32+353 = 385
- Now We have around 385 features to train our model 32 are continuous and 353 are from categorical features

- We have to treat the outliers because as I mentioned above for machine learning modelling it takes only normally distributed data with outliers in the data we tend lose the important information about rest of the data
- I used winsorization that is capping the anomalies or outliers to 5^{th} percentile for lower bound and 95^{th} percentile for upper bound then plotted the distribution

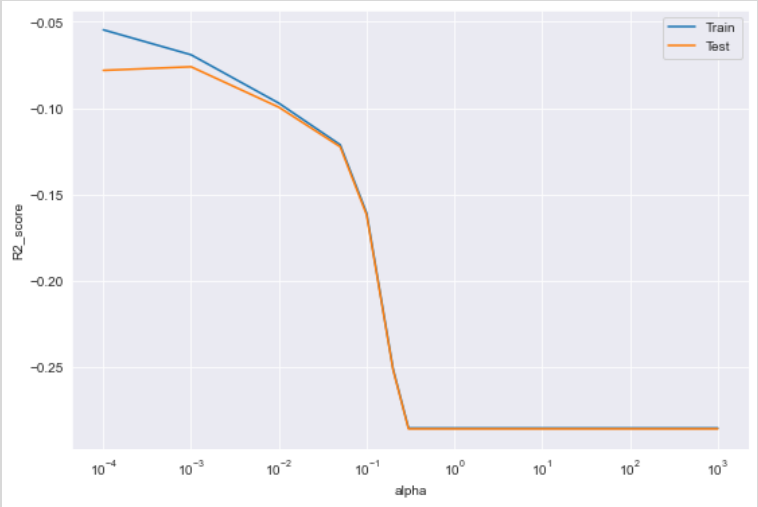
MODEL BUILDING

- We split our X_train Data to x_train and x_test in ratio 80:20 80% of data to train 20% for testing
- For modelling the x_train features should be of same scaling because we observed that many feature like lotfrontage which is in range from 50 to 200 and lot area in the range of 200,000 while modelling ,the model tend to get biased to high value feature we don't want that to happen so we bring to the same scale and compare.
- I used standardscale() which is also Z score regularization
- Now We are ready with the cleaned data we train our model
- Ridge Regression :
- For ridge we need to find the optimal lambda value for better result we use GridSearchcv as model selection we give lambda values from 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000 here first we start from 0.0001 then multiply with 10 then 5 then 0.1 and repeat it.
- We observed for lambda value 10 as best parameter
- after Evaluating we can see for lambda value approx. 10 the R2_score is high or peak

then decreases drastically



- For Lasso Regression :
- We do the steps involved in ridge , found out that the optimal value for Lasso is 0.001
- we can now plot R2 Score against different values of alpha for train and test sets
- For lamda value apprx 0.001 we can see the peak



- Now we if compare two models we can decide the most important feature as lasso make the slope of feature to zero we see what all columns it has rejected
- And in Ridge the coefficients that are closer to zero are irrelevant

Outputs of Both the Models :

- Created a table to represents the evaluation metrics :

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.94	0.92
R2 Score (Test)	0.93	0.93
RSS (Train)	8.53	11.29
RSS (Test)	2.87	2.92
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.09	0.10
RMSE (Test)	0.10	0.10

- Lasso has removed around 306 features from 384 columns or features rest is 79 features

RESULT

- Top 10 important features from 79 columns are in descending order from Ridge :
- **GrLivArea 1.09**
- **OverallQual_9 1.08**
- **OverallQual_8 1.08**
- **Neighborhood_Crawfor 1.08**
- **OverallCond_9 1.08**
- **Functional_Typ 1.07**
- **Exterior1st_BrkFace 1.07**
- **SaleCondition_Alloca 1.07**
- **CentralAir_Y 1.06**
- **TotalBsmtSF 1.05**

- Top 10 important features from lasso are in descending :
- **OverallQual_9 1.13**
- **GrLivArea 1.11**
- **OverallQual_8 1.11**
- **Neighborhood_Crawfor 1.09**
- **Exterior1st_BrkFace 1.08**
- **Functional_Typ 1.08**
- **CentralAir_Y 1.05**
- **Neighborhood_Somerst 1.04**
- **TotalBsmtSF 1.04**
- **Condition1_Norm 1.04**

CONCLUSION

- The variables significant in predicting the price of a house are:
- GrLivArea, OverallQual_9, OverallCond_9, OverallQual_8, Neighborhood_Crawfor, Functional_Typ, Exterior1st_BrkFace, SaleCondition_Alloca, CentralAir_Y, TotalBsmtSF, Neighborhood_Somerst, TotalBsmtSF and Condition1_Norm
- GrLivArea:an increase of 1 square foot of house area above ground, the price will increase by 1.09 to 1.11 times OverallQual_9 & OverallQual_8:
- if the overall material and finish of the house is Very Good or Excellent, the price of house will increase by 1.08 to 1.13 times
- Neighborhood_Crawfor: if Crawford is a nearby location, then the price of house will increase by 1.07 to 1.09 times
- Functional_Typ: if the home functionality is typical, then the price of house will increase by 1.07 to 1.08 times
- Exterior1st_BrkFace: the exterior covering on the house is Brick Face, the price of house will increase by 1.07 to 1.08 times.