ADVANCED HOUSE PRICE PREDICTION USING MACHINE LEARNING

TEAM PRESENTATION

Sushith Vanga	Srinidhi Muppidi	Yagnesh	Gamya	Vara Prasad	Amulya
CSU ID : 2830173	CSU ID : 2822455	CSU ID : 2825787	CSU ID :2829783	CSU ID : 2826343	CSU ID : 2830012

OUTLINE

- PROJECT SUMMARY
- TECHNOLOGY AND TOOLS
- DATA COLLECTION / PREPROCESSING
- FEATURE ENGINEERING
- MODEL BUILDING
- WEB APPLICATION

PROJECT SUMMARY

1

PROJECT SUMMARY

- The objective of this project is predict the house prices using machine learning techniques.
- Features like number of bedrooms,locality, year of bought etc are dependent variables and SalesPrice is the target feature.

INTRODUCTION

TRAINING THE MODEL

The supervised machine learning model is trained on the preprocessed data.

Feature selection, Feature engineering, model development, evaluation are part of this.

DEVELOPMENT OF WEB APPLICATION

Creating a Graphical user interface for this trained model.

Streamlit API

TECHNOLOGY AND TOOLS

2

TECHNOLOGY

There are 3 types of machine learning techniques

SUPERVISED LEARNING

The machine learns by using labeled data

Regression and Classification

Ex : cat or dog, spam or ham

UNSUPERVISED LEARNING

The machine is trained on unlabeled data

Association and Clustering

Ex : segmenting apples and bananas

REINFORCEMENT LEARNING

An agent interacts with its environment and learns by doing

Reward - based

Ex : Self driving cars

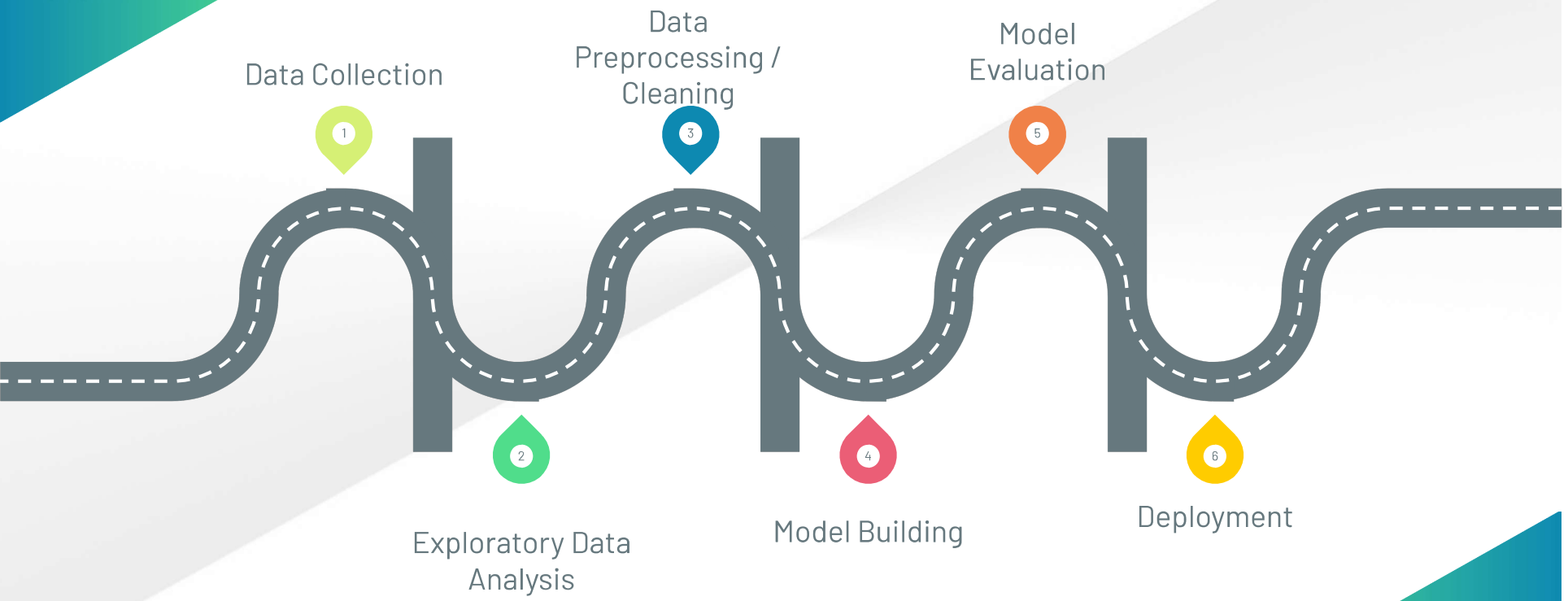
TOOLS

- Python
- Pycharm
- Pandas
- Numpy
- Sklearn
- Kaggle
- Streamlit

DATA COLLECTION / PROCESSING

3

DATA SCIENCE PROJECT LIFECYCLE



DATA COLLECTION

Data downloaded from **Kaggle Competitions**

- *Dataset prepared by Dean De Cock*

Variables :

Total of 79 variables present in the dataset

- *" SalePrice " represent final price at which the house was sold*
- *Remaining 78 variables represent different attributes like area, car parking, number of rooms etc*

DATA PREPROCESSING

Numerical Variables 38

- *Discrete variables - 17*
- *Continuous variables - 16*
- *Temporal variables - 4*

Categorical Variables - 43

FEATURE ENGINEERING

4

DATA PREPROCESSING

MISSING VALUES

Detection and correction

Numerical missing values are treated with median.

Categorical missing values are treated with “Missing” string.

OUTLIERS

Boxplot of all the dependent features .

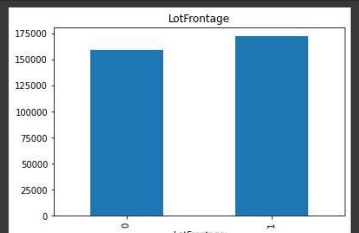
Dropping and replacing with mean, median and mode.

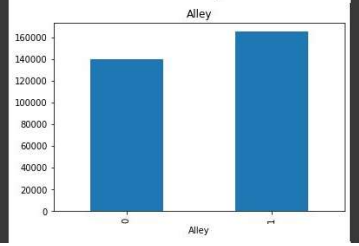
DATA TRANSFORMATION

Applied normalization technique on train and test data.

DATA PREPROCESSING (EDA)

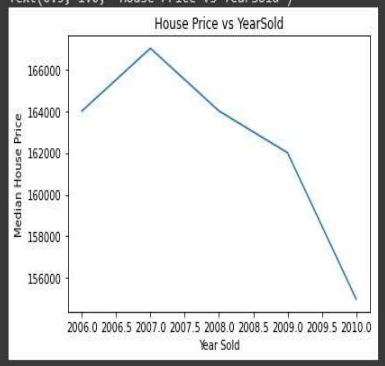
```
for feature in features_with_na:
    data = dataset.copy()

    # let's make a variable that indicates 1 if the observat
    data[feature] = np.where(data[feature].isnull(), 1, 0)

    # let's calculate the mean SalePrice where the informati
    data.groupby(feature)['SalePrice'].median().plot.bar()
    plt.title(feature)
    plt.show()
```





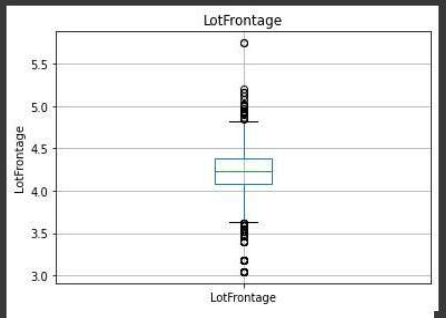

```
## Lets analyze the Temporal Datetime Variables
## We will check whether there is a relation between year the ho

dataset.groupby('YrSold')['SalePrice'].median().plot()
plt.xlabel('Year Sold')
plt.ylabel('Median House Price')
plt.title("House Price vs YearSold")
```

```
Text(0.5, 1.0, 'House Price vs YearSold')
```



Outliers

```
[ ] for feature in continuous_feature:
        data=dataset.copy()
        if 0 in data[feature].unique():
            pass
        else:
            data[feature]=np.log(data[feature])
            data.boxplot(column=feature)
            plt.ylabel(feature)
            plt.title(feature)
            plt.show()
```



MODEL BUILDING

5

MODEL BUILDING (DATA SPLITTING)

Validation Data

The sample of data used to provide an unbiased evaluation of a model fit on training dataset while tuning model.

20% - 25% of training data

Testing Data

The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

25% - 30% of entire data

Training Data

The actual dataset that we use to train the model..

70-75% of entire data.

MODEL BUILDING (REGRESSION ANALYSIS)

LINEAR REGRESSION

MAE - 0.138

MSE - 0.031

RMSE - 0.178

RMSE Logarithmic - 1.723

RIDGE REGRESSION

MAE - 0.030

MSE - 0.0018

RMSE - 0.043

RMSE Logarithmic - 3.141

KNN

MAE - 0.108

MSE - 0.0195

RMSE - 0.1397

RMSE Logarithmic - 1.967

RANDOM FOREST

MAE - 0.0113

MSE - 0.0002

RMSE - 0.0171

RMSE Logarithmic - 4.0647

XGBOOST

MAE - 0.1080

MSE - 0.0195

RMSE - 0.1397

RMSE Logarithmic - 1.9675

WEB APPLICATION

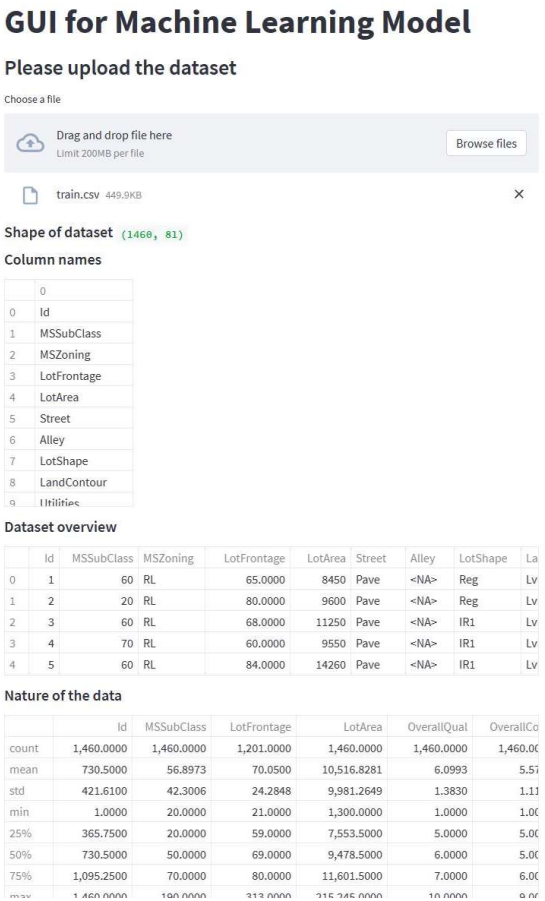
6

WEB APPLICATION

- Built GUI using **streamlit**.
- Integrating all the functionalities into the GUI.

MODEL BUILDING

- Option to upload csv files.
- Display of basic stats by default.
- Flexibility on feature engineering, model building and metrics.



Select the type of variable to analyse

Numerical variables

Number of numerical variables: 38

	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt
0	1	60	65.0000	8450	7	5	2003
1	2	20	80.0000	9600	6	8	1976
2	3	60	68.0000	11250	7	5	2001
3	4	70	60.0000	9550	7	5	1915
4	5	60	84.0000	14260	8	5	2000

Outlier detection

Plot out the outliers in the data

Finding the relation between categorical and dependent features

Plots between categorical and dependent?

Feature Engineering

Display the missing values

Do you want to correct the missing values?

Display missing values in numerical data and correct it?

Data Transformation

Select the data transformation technique to be implemented

Standardization

Split the data

Yes

x_train shape (978, 83)

x_test shape (482, 83)

y_train shape (978, 1)

y_test shape (482, 1)

Model Building

Select the machine learning model to train on

Select

Select

Linear Regression

Lasso Regression

K Nearest Neighbours

Random Forest

THANKS!