THE NORTHCAP UNIVERSITY

Project Report

House Price Prediction

Submitted By:

Rupali Taneja	18csu182
Saumya Achantani	18csu194
Saumya Gupta	18csu195

ACKNOWLEDGEMENT

This project has been done as part of our course Machine Learning at The Northcap University. Being extremely interested in everything having a relation to Machine Learning, this group project was a great occasion to give us the time to learn and confirm our interest in this field. In performing our project, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. The completion of this assignment gives us much Pleasure. We would like to show our gratitude to our teacher, **Ms. Poonam Chaudhary** for giving us a good guideline for the project throughout numerous consultations. Also, she gave us her valuable suggestions and ideas when we required them. She encouraged us to work on this project.

ABSTRACT

Real estate is the least transparent industry in our life. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various regression techniques and machine learning algorithms followed by neural networks in this pathway, and our results are not the sole determination of one technique rather it is the weighted mean of various techniques to give the most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. Also, we tried to make a recommendation system for providing an estimation of house prices based on the housing configurations provided by the user.

CONTENTS

1.OBJECTIVE

The main aim of this project is to predict the house price based on various features. Also, to practice feature engineering, machine learning, and neural network. Using different models in terms of minimizing the difference between predicted and actual values.

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges us to predict the final price of each home.

2. INTRODUCTION

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location. This project aims to predict house prices based on homes in Ames, Iowa, with regression analysis.

Fig 1 shows the overview of the project. These are the basic steps covered in this project such as Exploratory data analysis, regression, machine learning algorithms, etc.

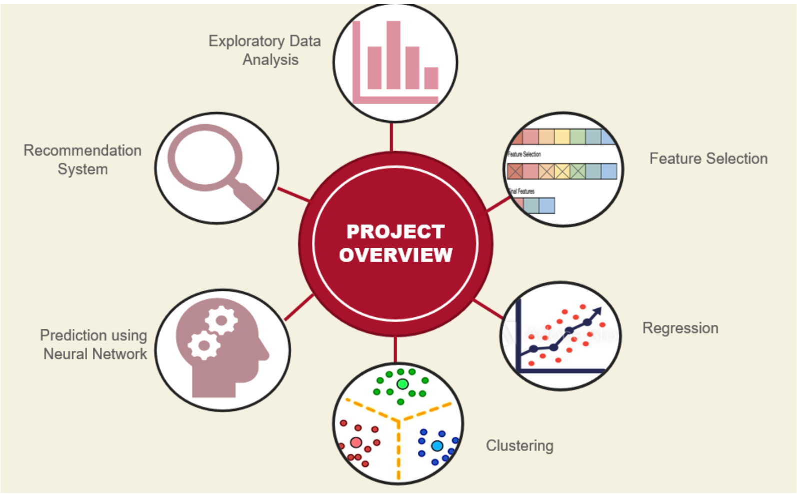


Fig-1 project overview

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses . Once we get a good fit, we will use this model to predict the monetary value of a house located at the area. A model like this would be very valuable for a real estate agent who could make use of the information provided on a daily basis.

3. DATASET DESCRIPTION

The data has been taken from the Kaggle competition named "House Prices: Advanced Regression Techniques". It contains *1460* training data points and 80 features that might help us predict the selling price of a house.

Here's a brief explanation of list of attributes.
SalePrice - the property's sale price in dollars. This is the target variable that we trying to predict.

- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet

- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

4. DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

4.1. Using python libraries

We're going to predict the SalePrice column let's start with it:

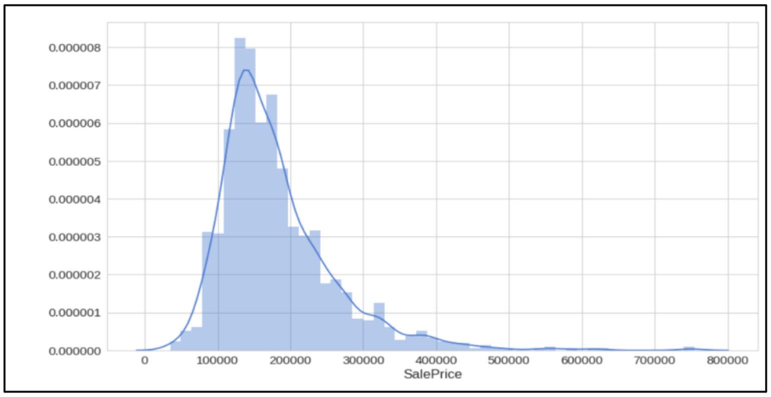


Fig 2: Sale Price distribution

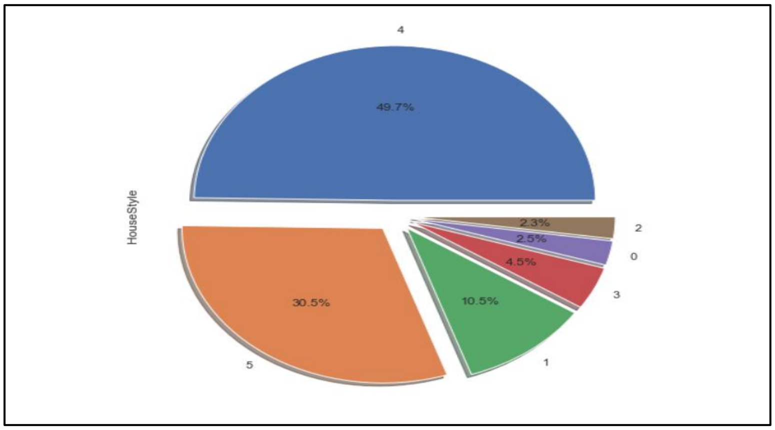


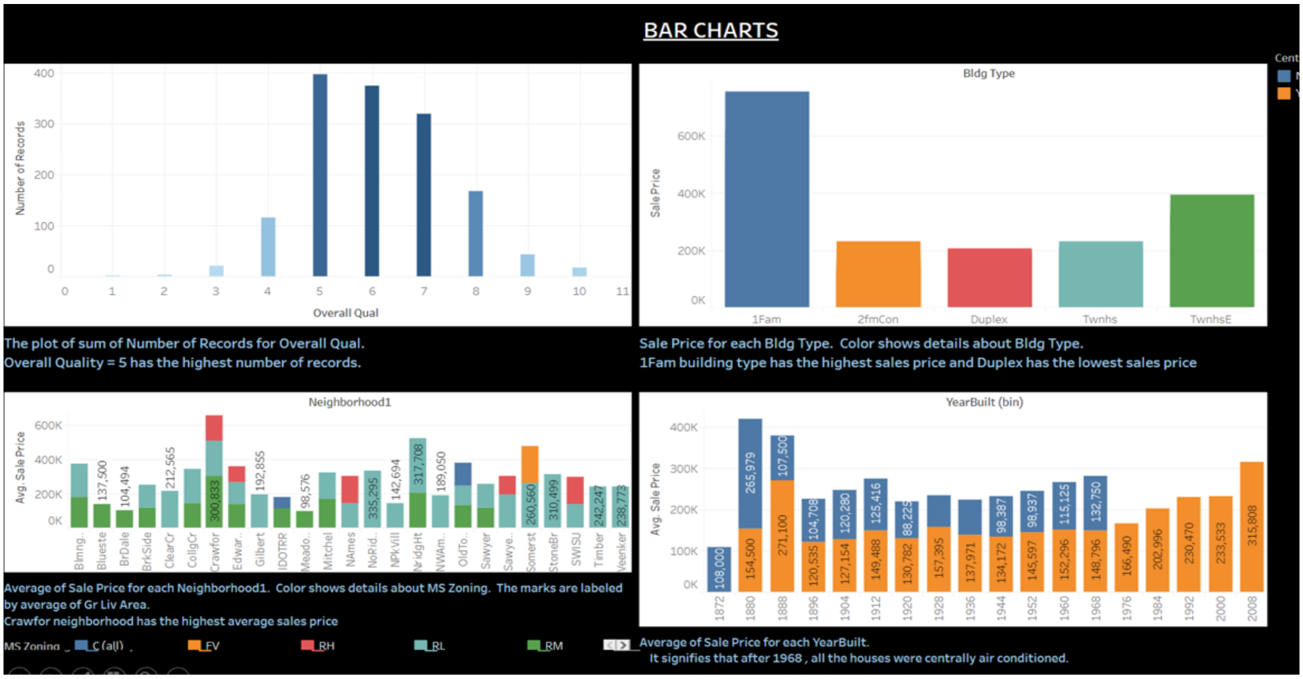
Fig 3:House Style Distribution

4.2. INTEGRATION OF TABLEAU INTO JUPYTER

To get better visualizations and hence better understanding of the data.

Steps to embed tableau workbook into Jupyter:

- To do this first create your Tableau workbook , then publish your workbook on the Tableau Public Server from the menu option.
- Once you have published the workbook, you can go to Tableau Public site using a browser and login. Now you can see your published Tableau workbook as shown below
- On the lower right hand corner of the published workbook, you can find a "Share" button.
- Once you client the share button, you can find the text box named Embed Code. Now, copy the text inside Embed Code.
- Go to your Jupyter notebook where you want the Tableau dashboard to be displayed.
- Enter into an empty code cell. Write %%HTML and then paste the copied text from Tableau public.


BAR CHARTS

The plot of sum of Number of Records for Overall Qual.
Overall Quality = 5 has the highest number of records.

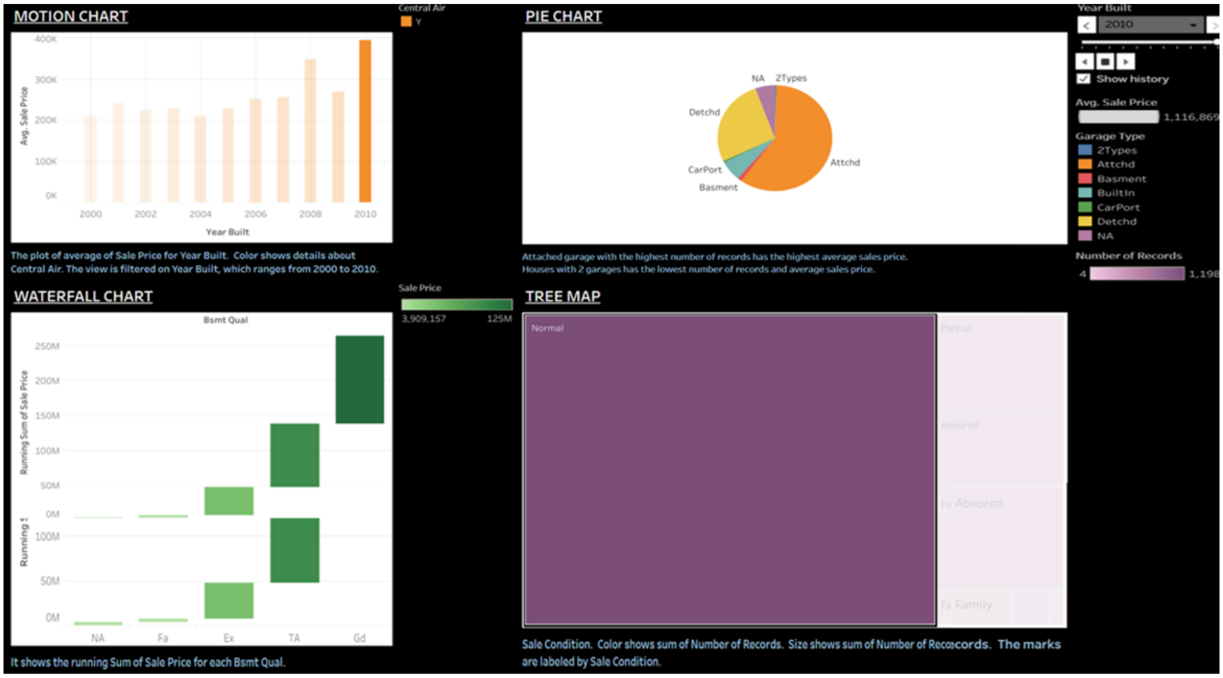
Sale Price for each Bldg Type. Color shows details about Bldg Type.
1Fam building type has the highest sales price and Duplex has the lowest sales price

Average of Sale Price for each Neighborhood1. Color shows details about MS Zoning. The marks are labeled by average of Gr Liv Area.
Crawfor neighborhood has the highest average sales price

MS Zoning 🟦 C (all) 🟧 FV 🟥 RH 🟩 RL 🟩 RM

Average of Sale Price for each YearBuilt.
It signifies that after 1968 , all the houses were centrally air conditioned.

Dashboard 1



MOTION CHART

The plot of average of Sale Price for Year Built. Color shows details about Central Air. The view is filtered on Year Built, which ranges from 2000 to 2010.

PIE CHART

Attached garage with the highest number of records has the highest average sales price. Houses with 2 garages has the lowest number of records and average sales price.

Year Built: 2010

☑ Show history

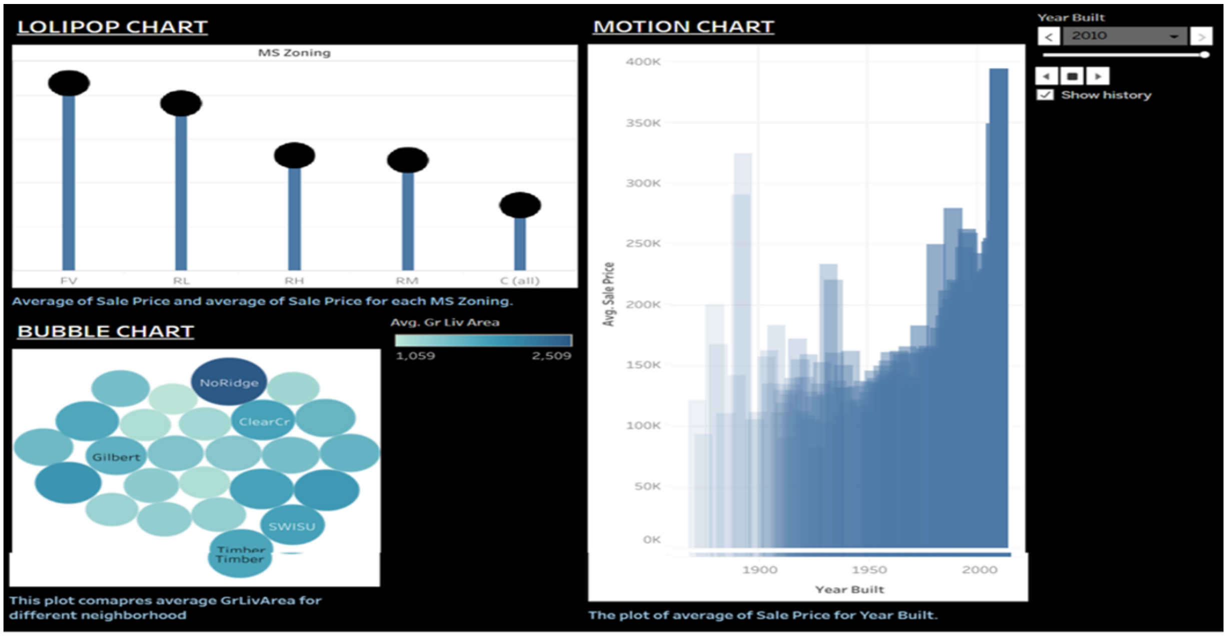
WATERFALL CHART

It shows the running Sum of Sale Price for each Bsmt Qual.

TREE MAP

Sale Condition. Color shows sum of Number of Records. Size shows sum of Number of Rececords. The marks are labeled by Sale Condition.

Dashboard 2



LOLIPOP CHART

Average of Sale Price and average of Sale Price for each MS Zoning.

BUBBLE CHART

This plot comapres average GrLivArea for different neighborhood

MOTION CHART

The plot of average of Sale Price for Year Built.

Year Built: 2010

☑ Show history

Dashboard 3

5. FEATURE SELECTION

- **Feature selection** is the process of reducing the number of input variables when developing a predictive model.
- It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

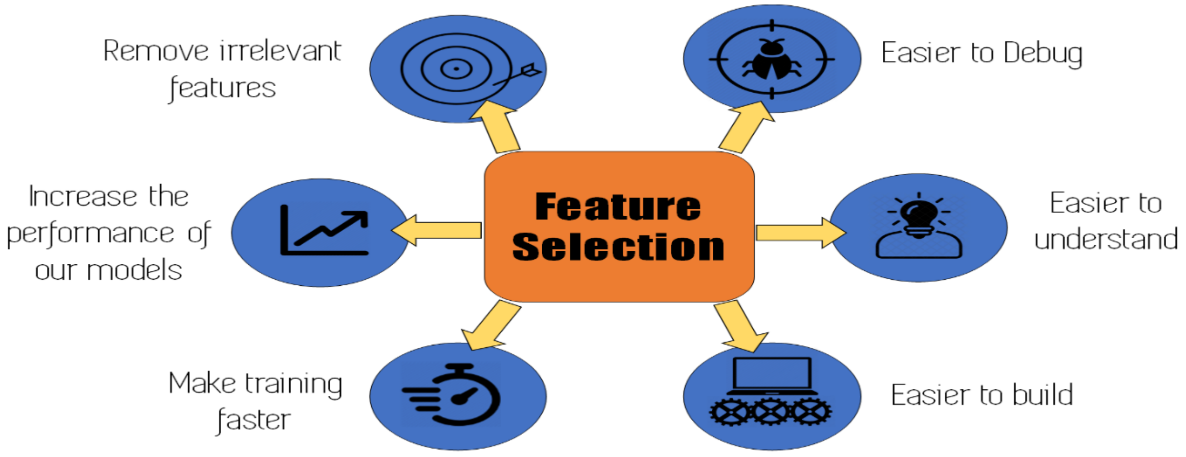


Fig 4:Importance of feature Selection

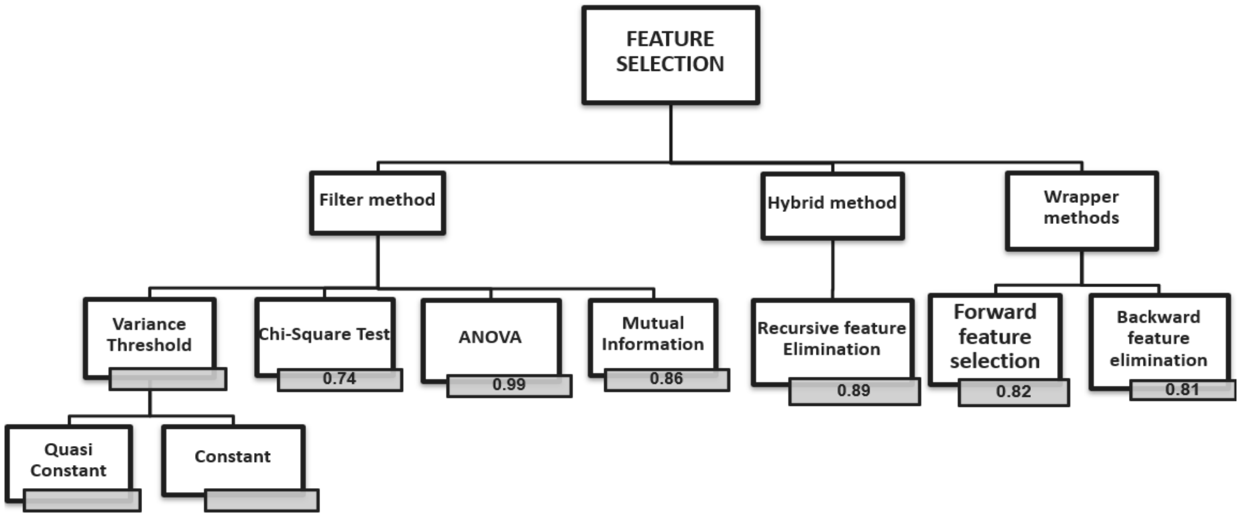


Fig 5:Feature selection methods

5.1. FILTER METHODS :

- **Filter methods** measure the relevance of features by their correlation with dependent variable.
- The selection of features is independent of any machine learning algorithms.
- Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable

5.1. Filter methods

5.1.1. Variance Threshold.

It is simple baseline approach to feature selection. It removes all features whose variance doesn’t meet some threshold.

It is of two types: Constant and Quasi constant

a) **Constant Features:** Features that show single values in all the observations in the dataset. These features provide no information that allows ML models to predict the target.

b) **Quasi Constant Features:** Features that are almost constant are known as quasi constant .These features have the same values for a very large subset of the outputs.

5.1.2. Anova.

Analysis **of** **Va**riance is a statistical method, used to check the means of two or more groups that are significantly different from each other. It assumes Hypothesis as

H0: Means of all groups are equal.

H1: At least one mean of the groups are different.

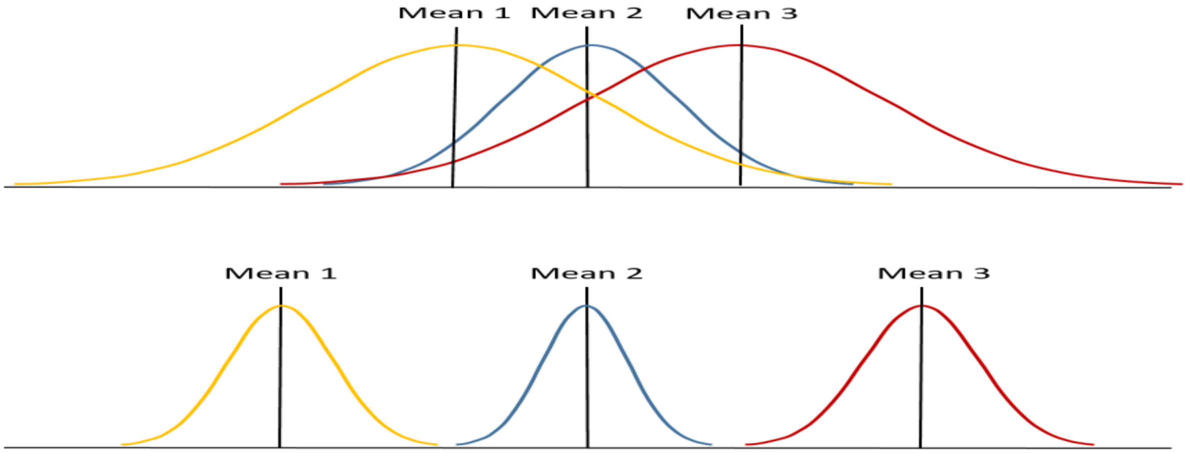


Fig 6: ANOVA

5.1.3. Chi Square Test.

$$X^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$

In feature selection, we aim to select the features which are highly dependent on the response.

When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value.

higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training

5.1.4. Mutual Information

Mutual information is a measure between two (possibly multi-dimensional) random variables X and Y, that quantifies the amount of information obtained about one random variable, through the other random variable.

5.2. Wrapper Methods

- In wrapper methods, the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset.
- It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.

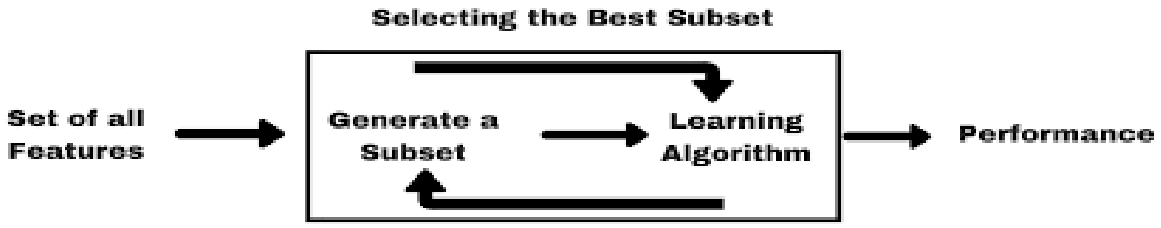


Fig 7:Wrapper approach

Wrapper methods applied:

5.2.1. Forward Feature Selection

Forward selection is an iterative method in which we start with having no feature in the model.

In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

5.2.2. Backward Feature Elimination

Backward elimination (or backward deletion) is the reverse process.

All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation

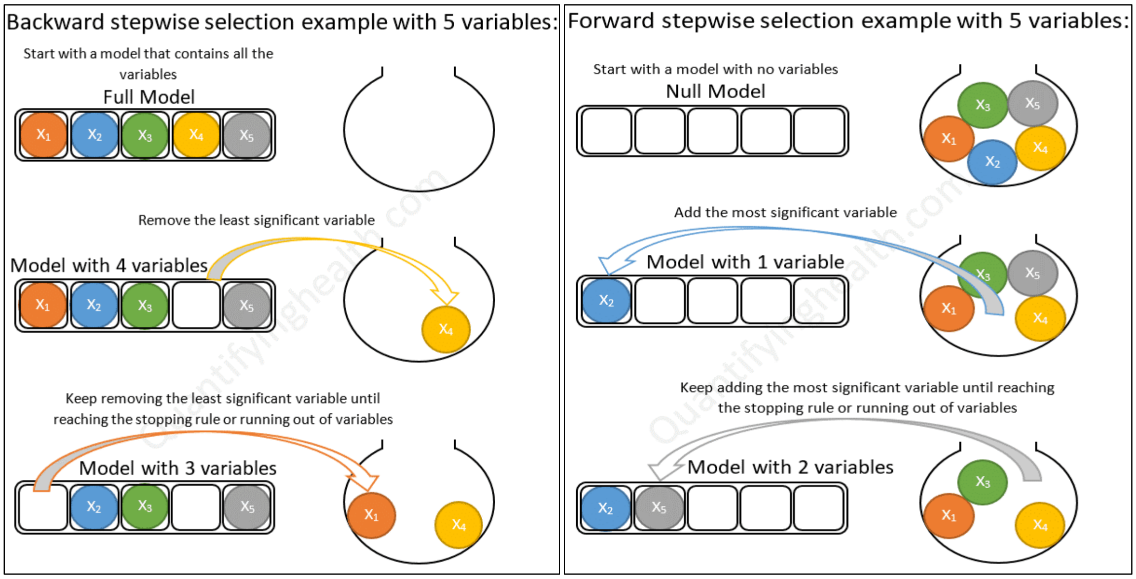

Backward stepwise selection example with 5 variables:
Start with a model that contains all the variables
Full Model
X_1 X_2 X_3 X_4 X_5
Remove the least significant variable
Model with 4 variables
X_1 X_2 X_3 X_5
X_4
Keep removing the least significant variable until reaching the stopping rule or running out of variables
Model with 3 variables
X_2 X_3 X_5
X_1 X_4
Forward stepwise selection example with 5 variables:
Start with a model with no variables
Null Model
X_3 X_5 X_1 X_2 X_4
Add the most significant variable
Model with 1 variable
X_2
X_3 X_5 X_1 X_4
Keep adding the most significant variable until reaching the stopping rule or running out of variables
Model with 2 variables
X_2 X_5
X_3 X_1 X_4

Fig 8:Backward and Forward Feature Selection

5.3. Hybrid Methods

Hybrid methods combine the different approaches to get the best possible feature subset.

For example, start by performing filter methods by eliminating constant, quasi-constant and duplicated features. Then, in the second step, you could use wrapper methods to select the best feature subset from the previous step. This is just one simple, high-level approach.

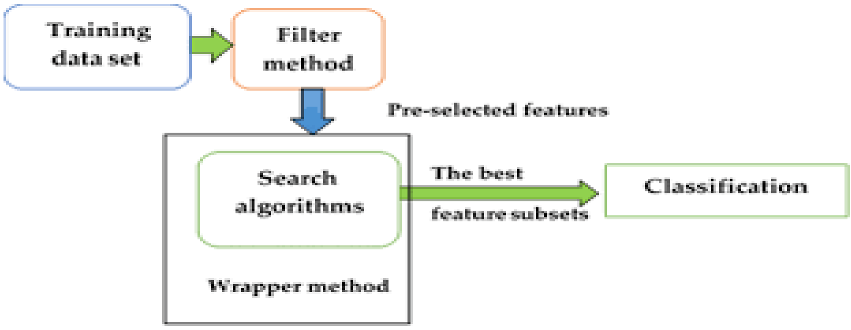


Fig 9:Hybrid Approach

Hybrid Methods applied:

5.3.1 Recursive Feature Elimination :

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid.

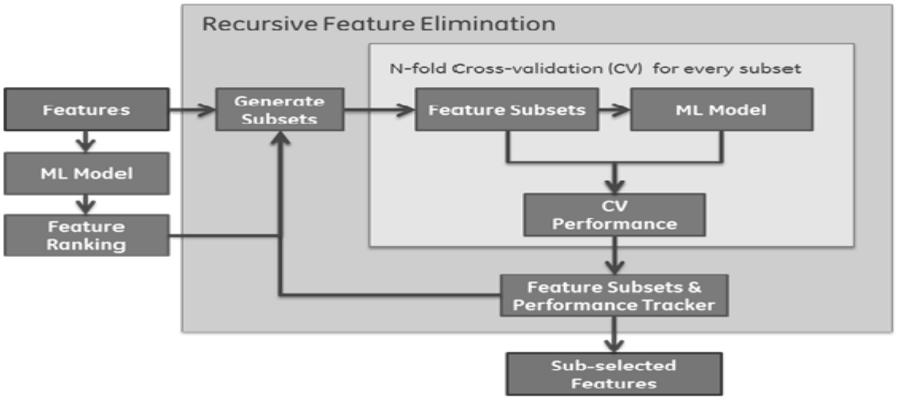

Recursive Feature Elimination
N-fold Cross-validation (CV) for every subset
Features
ML Model
Feature Ranking
Generate Subsets
Feature Subsets
ML Model
CV Performance
Feature Subsets & Performance Tracker
Sub-selected Features

Fig 10:Recursive feature elimination

RESULT :

RECURSIVE FEATURE ELIMINATION worked the best for this dataset

6. PREDICTING THE SALE PRICE

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the casual relationship effect between the variables.

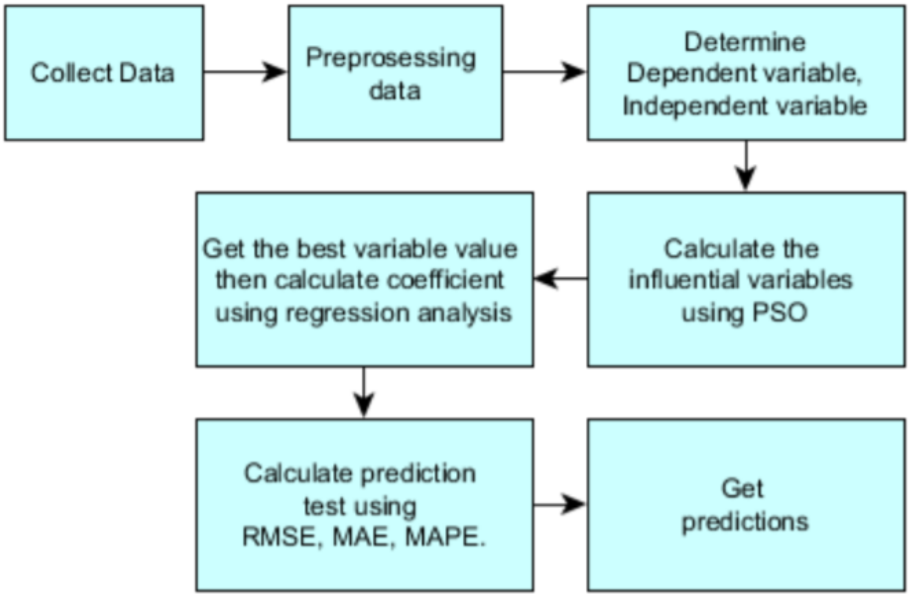

Collect Data
Preprosessing data
Determine Dependent variable, Independent variable
Calculate the influential variables using PSO
Get the best variable value then calculate coefficient using regression analysis
Calculate prediction test using RMSE, MAE, MAPE.
Get predictions

Fig 11. Flow Diagram for Regression Models

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. I'll explain this in more details in coming sections.

Why do we use Regression Analysis?

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

We have used various regression models to train our data and get the best results.

6.1. Linear Regression

- Linear regression models assume that the relationship between a dependent continuous variable Y and one or more explanatory (independent) variables X is linear (that is, a straight line).
- It's used to predict values within a continuous range (e.g. sales, price)

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).
For training the Linear Regression Model, we have considered independent variable as GrLivArea as it has highest correlation with the SalePrice. Taking it into consideration, below is the graph of best fit line obtained.

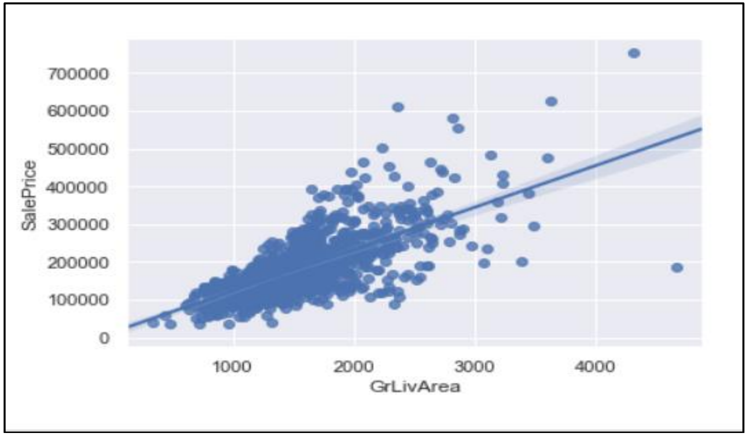


Fig 12. Best fit line

We can evaluate the model performance using the metric R-square.
The R-squared value for Linear Regression model on our dataset came out be 0.44 .

6.2. Polynomial Regression

- Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.
- Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x)

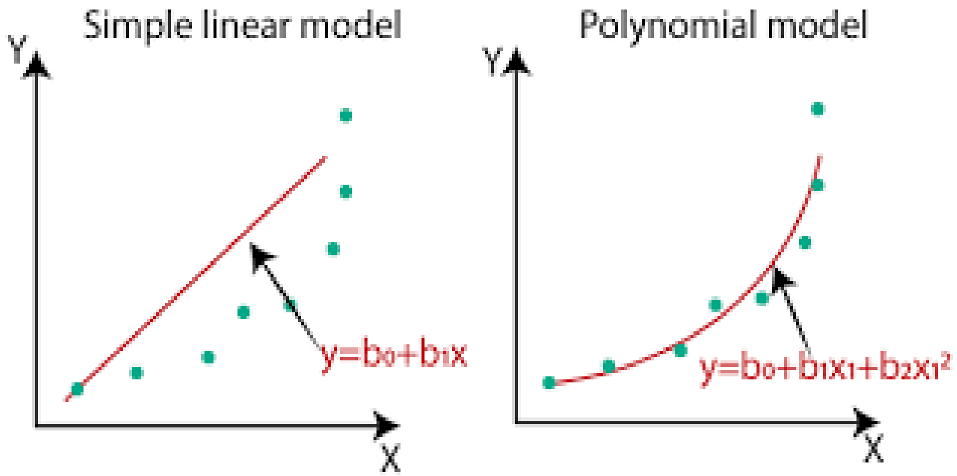


Fig 13. Linear regression vs Polynomial regression

This regression considers the nth degree of independent variable and hence avoids underfitting in some cases.
In our case, we have considered 4th degree of the independent variable and the R square value came out to be 0.50 .

6.3. Multiple Regression

- A more complex, multi-variable linear equation might look like this, where w represents the coefficients or weights, our model will try to learn.

$$Y(x_1, x_2, x_3) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$$

The variables x1, x2, x3 represent the attributes or distinct pieces of information, we have about each observation.

6.4. Ridge Regression

- Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

- Above, we saw the equation for linear regression.It can be represented as:

  y=a+ b*x

- This equation also has an error term. The complete equation becomes:

  y=a+b*x+e (error term) [error term is the value needed to correct for a prediction error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the **bias** and second is due to the **variance**. Prediction error can occur due to any one of these two or both components. Here, we'll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda).

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

In this equation, we have two components. First one is least square term and other one is lambda of the summation of β2 (beta- square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

After prediction, the predicted R square value is 0.73 which is highest among all above regression models.

6.5. XGBoost Regression

- XGBoost is an implementation of gradient boosted decision trees designed for speed and performance
- XGBoost stands for eXtremeGradientBoosting.
- Gradient boosting is a method where the new models are created that computes the error in the previous model and then leftover is added to make the final prediction. It uses a gradient descent algorithm that is the reason it is called a "Gradient Boosting Algorithm".

The reason behind the good performance of XGboost -

1. *Regularization:*

   This is considered to be as a dominant factor of the algorithm. Regularization is a technique that is used to get rid of overfitting of the model.

2. Cross-Validation:

   We use cross-validation by importing the function from sklearn but XGboost is enabled with inbuilt CV function.

3. Missing Value:

   It is designed in such a way that it can handle missing values. It finds out the trends in the missing values and apprehends them.

4. Flexibility:

   It gives the support to objective functions. They are the function used to evaluate the performance of the model and also it can handle the user-defined validation metrics.

5. Save and load:

   It gives the power to save the data matrix and reload afterwards that saves the resources and time.

6.6. Support Vector regression

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

In support Vector Regression, Our main aim here is to decide a decision boundary at ‘a’ distance from the original hyperplane such that data points closest to the hyperplane or the support vectors are within that boundary line.

Regression conclusion plot:

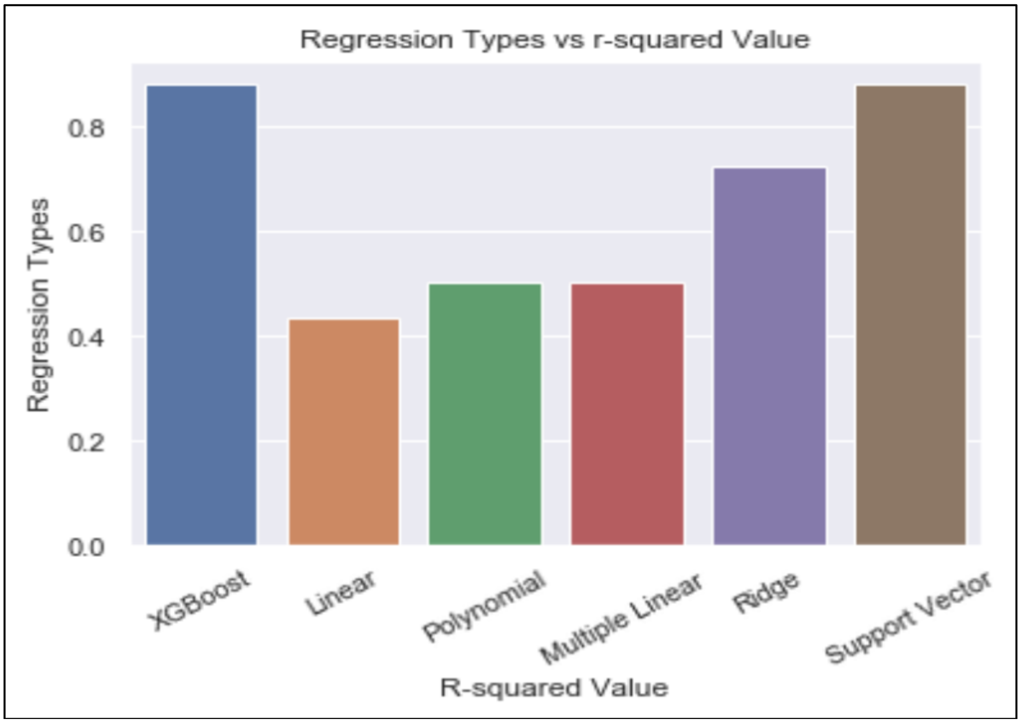


Fig 14. Comparison of R squared values of different regression.

After applying various regression models to predict Sales Price, the highest R Square value achieved by using XG-Boost and SVR.

7. NEURAL NETWORKS

- A neural network is an artificial neural network with one or more layers between the input and output layers.
- **Neural networks** offer a number of **advantages**, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training
- Neural networks work better at predictive analytics because of the hidden layers. Linear regression models use only input and output nodes to make predictions. Neural network also use the hidden layer to make predictions more accurate.

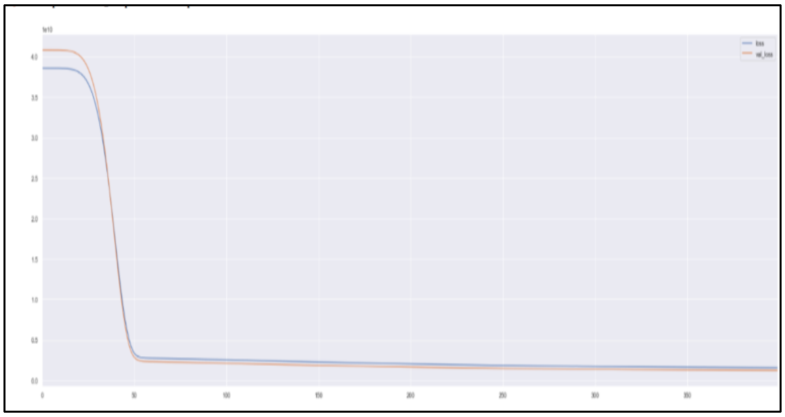


Fig 15:

The following graph between no of epochs and mean squared error shows loss function for training and validation sets

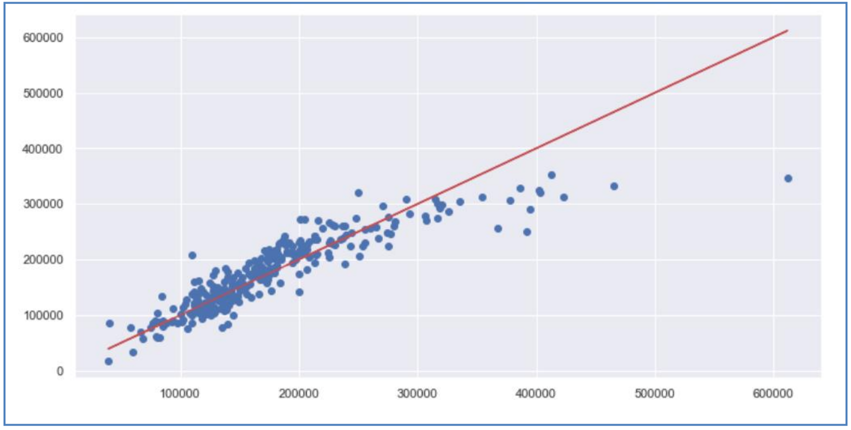


Fig 16:

- Library used - Tensorflow
- Activation Function used is ReLU
- No of epoch = 400
- No of neurons in each layer (except output)= 24
- Predicted R2 Square 0.79

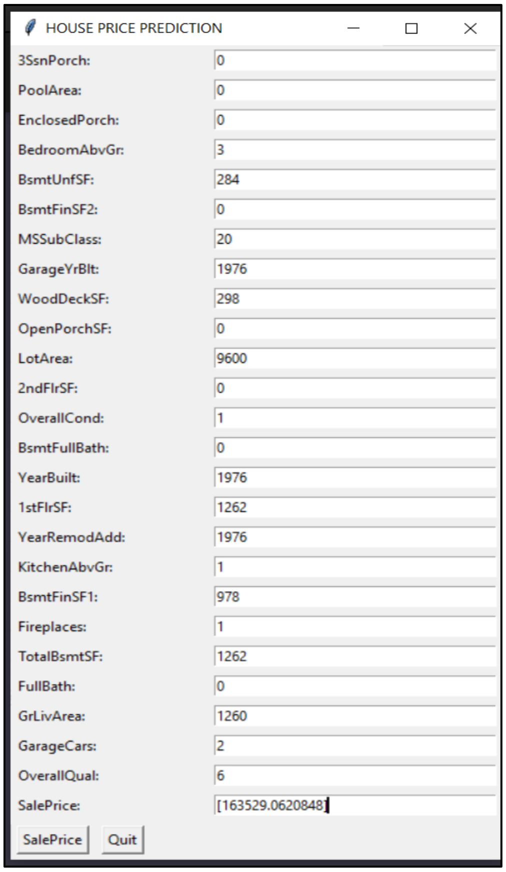
8. GUI PROGRAMMING USING TKINTER

HOW TO MAKE THIS PROJECT MORE USERS FRIENDLY?

- Python with tkinter is the fastest and easiest way to create the GUI applications. Creating a GUI using tkinter is an easy task.
- The big advantage of using this is that it's cross platform - so that it will work fine on Linux, Windows and Mac OS without any further installations.

To create a tkinter app:

1. Importing the module – tkinter
2. Create the main window (container)
3. Add any number of widgets to the main window
4. Apply the event Trigger on the widgets.



HOUSE PRICE PREDICTION

3SsnPorch:	0
PoolArea:	0
EnclosedPorch:	0
BedroomAbvGr:	3
BsmtUnfSF:	284
BsmtFinSF2:	0
MSSubClass:	20
GarageYrBlt:	1976
WoodDeckSF:	298
OpenPorchSF:	0
LotArea:	9600
2ndFlrSF:	0
OverallCond:	1
BsmtFullBath:	0
YearBuilt:	1976
1stFlrSF:	1262
YearRemodAdd:	1976
KitchenAbvGr:	1
BsmtFinSF1:	978
Fireplaces:	1
TotalBsmtSF:	1262
FullBath:	0
GrLivArea:	1260
GarageCars:	2
OverallQual:	6
SalePrice:	[163529.0620848]

Fig 17.Prediction interface

This provides user to enter the house configurations and provide them the value of estimated sale price. This will help user to select appropriate house in their budget.

9. CLASSIFICATION

Classification is a process of categorizing a given set of data into classes. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

Since classification is a type of supervised learning, even the targets are also provided with the input data.

How is classification used in this project?

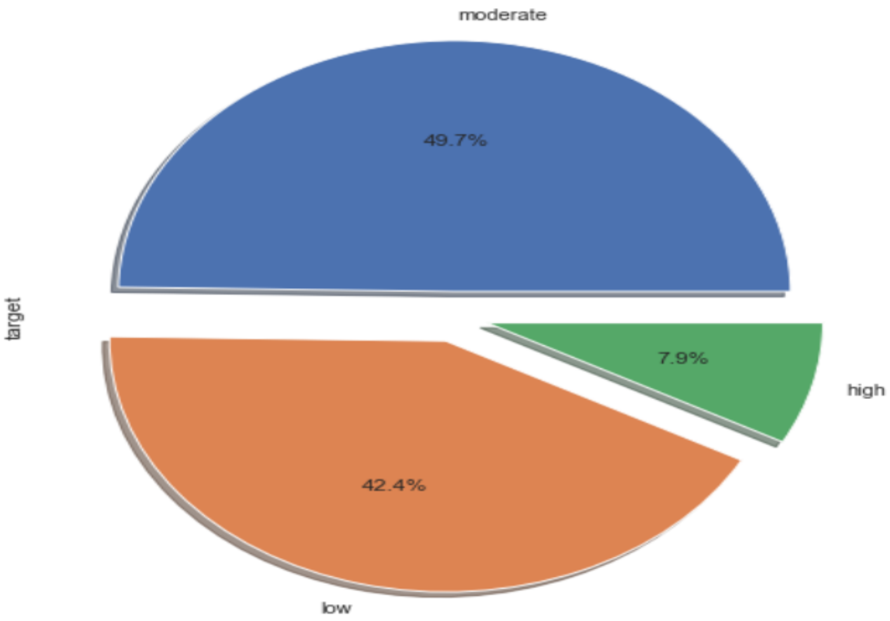


Fig 18. Categorising target variable sales price into 3 categories

- ✓ Low
- ✓ Moderate
- ✓ High

Classification Terminologies in Machine Learning :

- **Classifier** – It is an algorithm that is used to map the input data to a specific category.
- **Classification Model** – The model predicts or draws a conclusion to the input data given for training, it will predict the class or category for the data.
- **Feature** – A feature is an individual measurable property of the phenomenon being observed.
- **Binary Classification** – It is a type of classification with two outcomes, for eg – either true or false.
- **Multi-Class Classification** – The classification with more than two classes, in multi-class classification each sample is assigned to one and only one label or target.
- **Multi-label Classification** – This is a type of classification where each sample is assigned to a set of labels or targets.
- **Initialize** – It is to assign the classifier to be used for the
- **Train the Classifier** – Each classifier in sci-kit learn uses the fit(X, y) method to fit the model for training the train X and train label y.
- **Predict the Target** – For an unlabeled observation X, the predict(X) method returns predicted label y.
- **Evaluate** – This basically means the evaluation of the model i.e classification report, accuracy score, etc.Support Vector Machine

Classification models used-

1.Support Vector Classification

The support vector machine is a classifier that represents the **training data as points in space** separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.

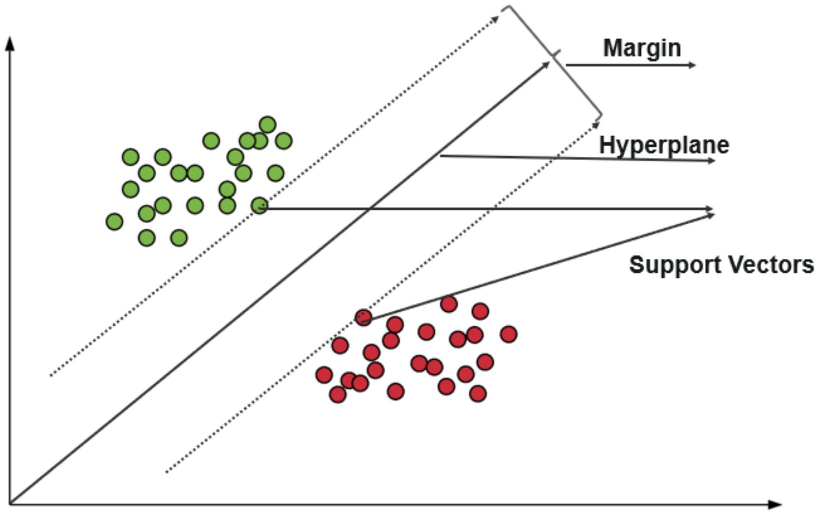


Fig 18. Diagram of Support Vector Classification

Advantages and Disadvantages

It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces.

The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

Use cases

- Business applications for comparing the performance of a stock over a period of time
- Investment suggestions
- Classification of applications requiring accuracy and efficiency

2.K-Nearest Neighbor

It is a lazy learning algorithm that **stores all instances corresponding to training data in n-dimensional space**. It is a **lazy learning algorithm** as it does not focus on constructing a general internal model, instead, it works on storing instances of training data.

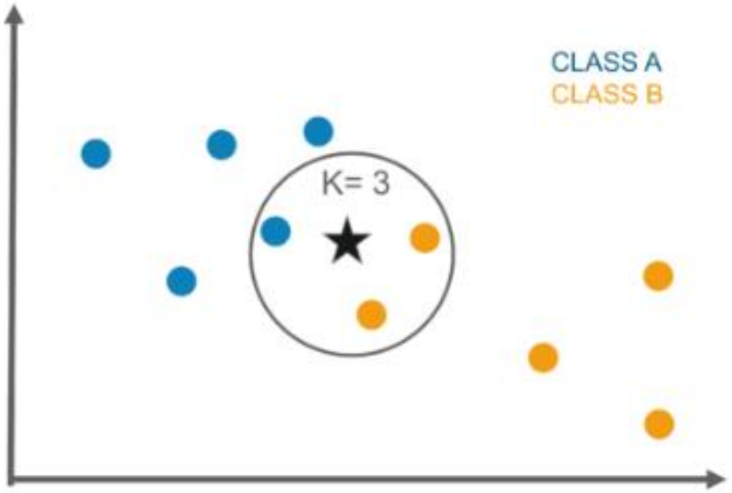


Fig 19. KNN Explanatory diagram

Classification is computed from a simple majority vote of the k nearest neighbors of each point. It is supervised and takes a bunch of labeled points and uses them to label other points. To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbors. It has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point. The “k” is the number of neighbors it checks.

Advantages And Disadvantages

This algorithm is quite simple in its implementation and is robust to noisy training data. Even if the training data is large, it is quite efficient.

The only disadvantage with the KNN algorithm is that there is no need to determine the value of K and computation cost is pretty high compared to other algorithms.

Use Cases

- Industrial applications to look for similar tasks in comparison to others
- Handwriting detection applications
- Image recognition
- Video recognition
- Stock analysis

Classifier Evaluation

The most important part after the completion of any classifier is the evaluation to check its accuracy and efficiency. There are a lot of ways in which we can evaluate a classifier. Let us take a look at these methods listed below.

- **Holdout Method**

This is the most common method to evaluate a classifier. In this method, the given data set is divided into two parts as a test and train set 20% and 80% respectively.

The train set is used to train the data and the unseen test set is used to test its predictive power.

- **Cross-Validation**

Over-fitting is the most common problem prevalent in most of the machine learning models. K-fold cross-validation can be conducted to verify if the model is over-fitted at all.

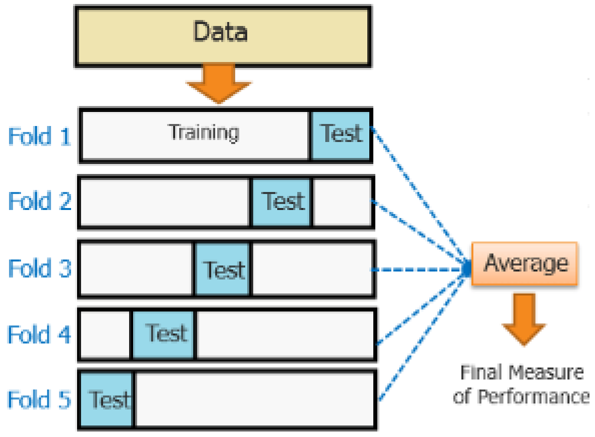


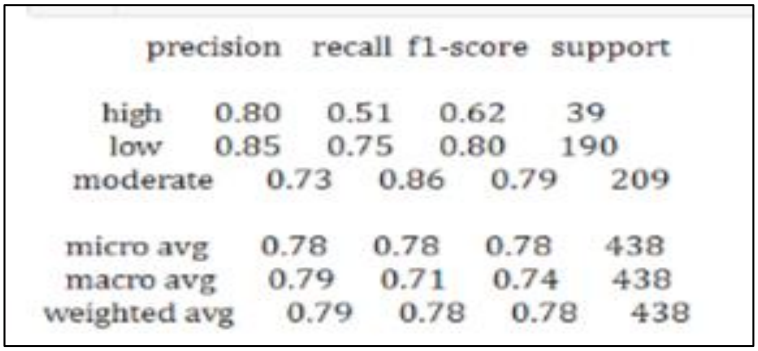
Fig 20.Explanatory diagram of cross validation

In this method, the data set is randomly partitioned into **k mutually exclusive** subsets, each of which is of the same size. Out of these, one is kept for testing and others are used to train the model. The same process takes place for all k folds.

- **Classification Report**

A classification report will give the following results:-

a)KNN

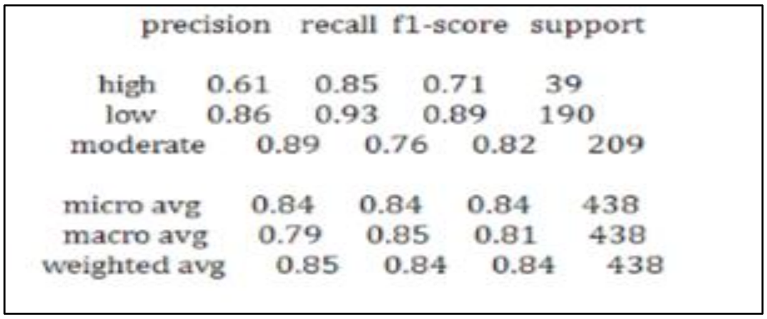


```
              precision    recall  f1-score   support

        high       0.80      0.51      0.62        39
         low       0.85      0.75      0.80       190
    moderate       0.73      0.86      0.79       209

   micro avg       0.78      0.78      0.78       438
   macro avg       0.79      0.71      0.74       438
weighted avg       0.79      0.78      0.78       438
```

b)Naïve Bayes

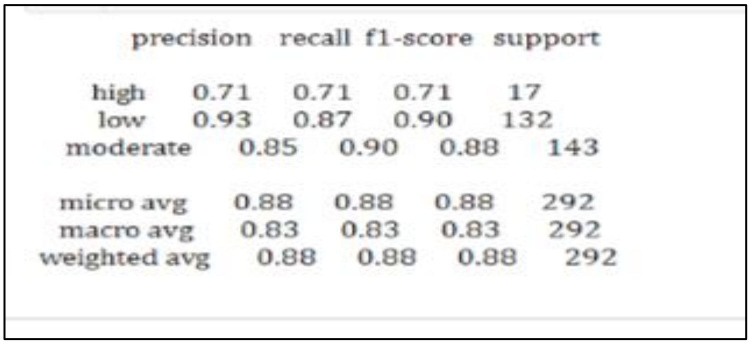


```
              precision    recall  f1-score   support

        high       0.61      0.85      0.71        39
         low       0.86      0.93      0.89       190
    moderate       0.89      0.76      0.82       209

   micro avg       0.84      0.84      0.84       438
   macro avg       0.79      0.85      0.81       438
weighted avg       0.85      0.84      0.84       438
```

c)SVM



```
              precision    recall  f1-score   support

        high       0.71      0.71      0.71        17
         low       0.93      0.87      0.90       132
    moderate       0.85      0.90      0.88       143

   micro avg       0.88      0.88      0.88       292
   macro avg       0.83      0.83      0.83       292
weighted avg       0.88      0.88      0.88       292
```

- **Accuracy**
  - Accuracy is a ratio of correctly predicted observation to the total observations
  - True Positive: The number of correct predictions that the occurrence is positive.
  - True Negative: Number of correct predictions that the occurrence is negative.
- **F1- Score**
  - It is the weighted average of precision and recall
- **Precision And Recall**
  - Precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that have been retrieved over the total number of instances. They are basically used as the measure of relevance.

ROC Curve

Receiver operating characteristics or ROC curve is used for visual comparison of classification models, which shows the relationship between the true positive rate and the false positive rate. The area under the ROC curve is the measure of the accuracy of the model.

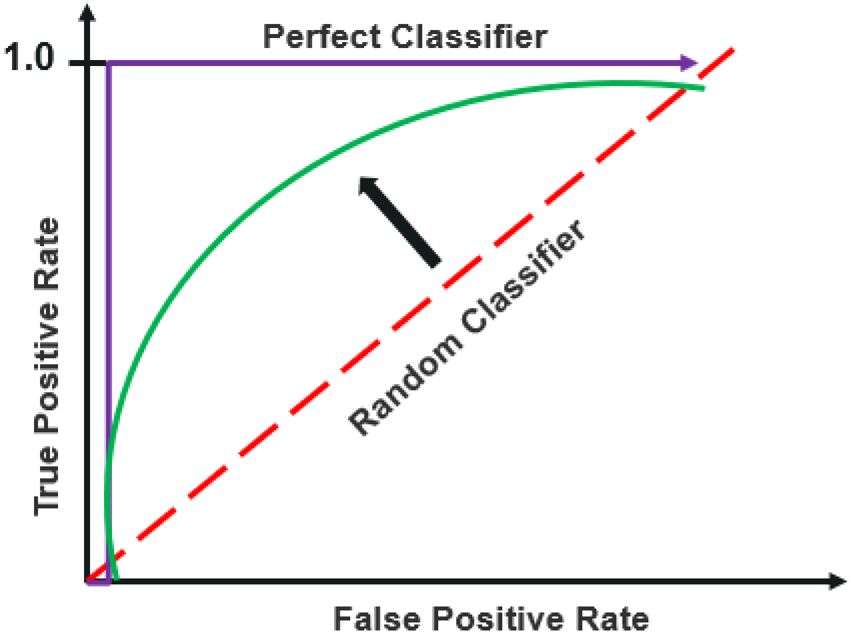


Fig 21. ROC Explanation diagram

Conclusion Report of Classification:

Models	Accuracy
KNN	78%
NAÏVE BAYES	84%
SVM	90%

10. SUMMARY

- By using XGBoost regression model we are able to predict sales price with 88% accuracy**.**
- We have covered regression, neural networks, classification, GUI in this project.
- Predicting house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well.
- In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

11. BIBLIOGRAPHY

- https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- https://medium.com/analytics-vidhya/predicting-house-prices-using-classical-machine-learning-and-deep-learning-techniques-ad4e55945e2d
- https://thesai.org/Downloads/Volume8No10/Paper_42-Modeling_House_Price_Prediction_using_Linear_Regression.pdf
- https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1
- https://github.com/ShivaniMangal/House-Price-Prediction-EDA-and-ML-