Predicting house prices using ML

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.

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

In the contemporary digital era, useful information of the society can be retrieved from a wide variety of sources and stored in the form of structured, unstructured and semi-structured formats. In the analysis of economic phenomena or social observations, advancement of innovative technology makes it possible to systematically extract the relevant information, transform them into complex data formats and structures, and then perform suitable analyses. Because of these new circumstances, traditional data processing and analytical tools may not be able to capture, process and analyse highly complex

information in the social and economic worlds. New techniques have been developed in response to the treatment of the colossal amount of available data.

Machine learning is one of the cutting edge techniques that can be used to identify, interpret, and analyse hugely complicated data structures and patterns . It allows consequential learning and improves model predictions with a systematic input of more recent data …. Contemporary research on machine learning is a subset of artificial intelligence (AI) that attempts to train computers with new knowledge through input of data, such as texts, images, numerical values, and so on, and support its interaction with other computer networks. Machine learning is about 'the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions'.

Machine learning can be broadly categorised into three types, namely supervised learning, unsupervised learning and semi-supervised learning. In essence, supervised machine learning algorithms intend to identify a function that is able to give accurate out-of-sample forecasts. For example, in property research, if an investigator intends to make forecast of housing prices yi from its physical, neighbourhood and accessibility characteristics xijfrom a sample of *n* apartments, one can assume L(yiˆ,yi) to be the prediction loss function. A machine learning algorithm will look for a function fˆ that produces lowest expected prediction loss E(yi,xij)[L(fˆ(xij),yi]on the *test* data from the same distribution…. A model trained through supervised learning is said to be successful if it can make predictions within an acceptable level of accuracy. Examples of supervised learning include linear regression and support vector machine.

- **Creative feature engineering**

Advanced regression techniques like random forest and gradient boosting.

- **Fun with Real Estate Data**

Use Rmarkdown to learn advanced regression techniques like random forests and XGBoost

- **XGBoost with Parameter Tuning**

Implement LASSO regression to avoid multicollinearity

Includes linear regression, random forest, and XGBoost models as well

- **Ensemble Modeling: Stack Model Example**

Use “ensembling” to combine the predictions of several models

Includes GBM (gradient boosting machine), XGBoost.

- ## A Clear Example of Overfitting

Learn about the dreaded consequences of overfitting data

Comprehensive Data Exploration with Python

Understand how variables are distributed and how they interact

Apply different transformations before training machine learning models

- ## House Prices EDA

Learn to use visualization techniques to study missing data and distributions

Includes correlation heatmaps, pairplots, and t-SNE to help inform appropriate inputs to a linear model

A Study on Regression Applied to the Ames Dataset

Demonstrate effective tactics for feature engineering

Explore linear regression with different regularization methods including ridge, LASSO, and ElasticNet using scikit-learn

Regularized Linear Models

Build a basic linear model

Try more advanced algorithms including XGBoost

```
import numpy as np
import pandas as pd

%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
color = sns.color_palette()
sns.set_style('darkgrid')
import warnings
def ignore_warn(*args, **kwargs):
```

3

```
    pass
warnings.warn = ignore_warn

from scipy import stats
from scipy.stats import norm, skew


pd.set_option('display.float_format', lambda x: '{:.3f}'.format(x))


from subprocess import check_output
print(check_output(["ls", "../input"]).decode("utf8"))

train = pd.read_csv('/kaggle/input/house-prices-advanced-regression-techniques/train.csv')
test = pd.read_csv('/kaggle/input/house-prices-advanced-regression-techniques/test.csv')
train.head(5)
```

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice	
0	1	60	RL	65.000	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.000	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice	
2	3	60	RL	68.0000	112250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0000	95500	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0000	142260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000

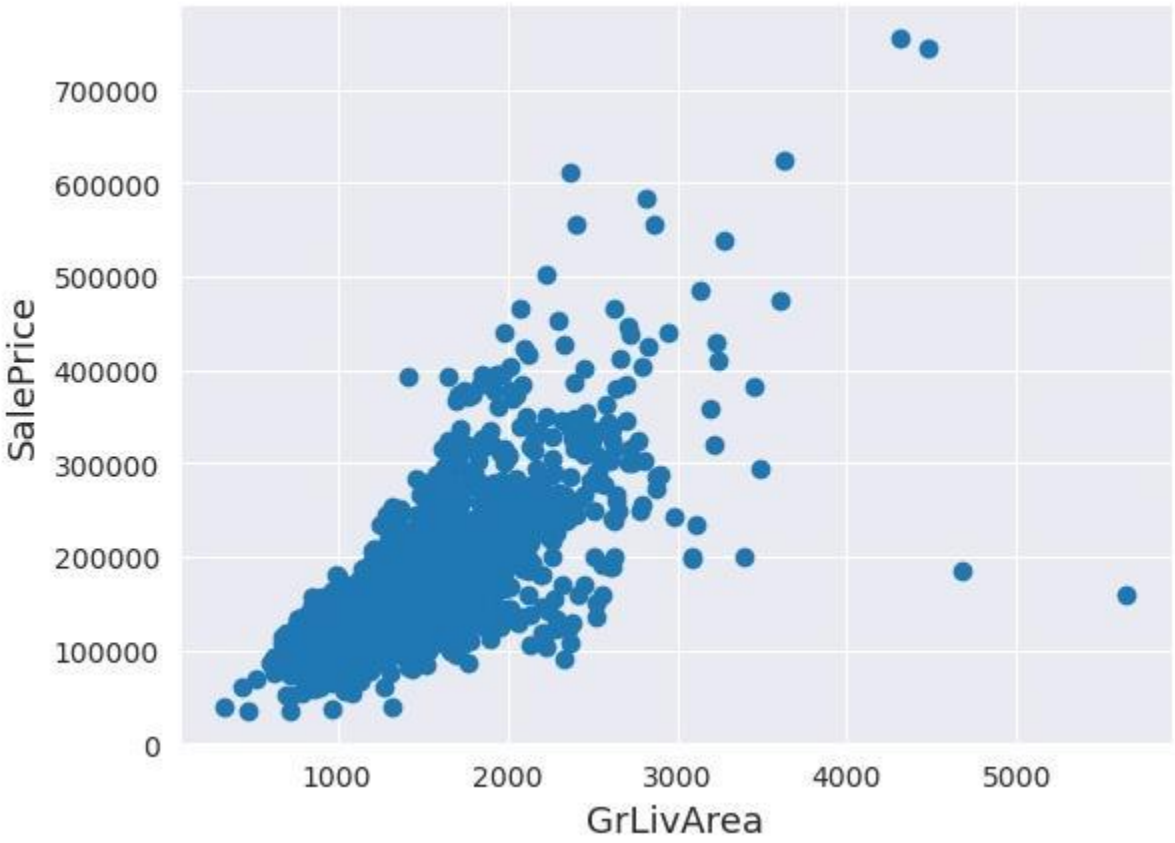
```
test.head(5)
```

5

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	
0	1461	20	RH	80.0000	11622	Pave	NaN	Reg	Lvl	AllPub	...	120	0	NaN	MnPrv	NaN	0	6	2010	WD	Normal
1	1462	20	RL	81.0000	14267	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN	Gar2	12500	6	2010	WD	Normal
2	1463	60	RL	74.0000	13830	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	MnPrv	NaN	0	3	2010	WD	Normal
3	1464	60	RL	78.0000	9978	Pave	NaN	IR1	Lvl	AllPub	...	0	0	NaN	NaN	NaN	0	6	2010	WD	Normal
4	1465	120	RL	43.0000	50005	Pave	NaN	IR1	HLS	AllPub	...	144	0	NaN	NaN	NaN	0	1	2010	WD	Normal

Outliers

```
fig, ax = plt.subplots()
ax.scatter(x = train['GrLivArea'], y = train['SalePrice'])
plt.ylabel('SalePrice', fontsize=13)
plt.xlabel('GrLivArea', fontsize=13)
plt.show()
```



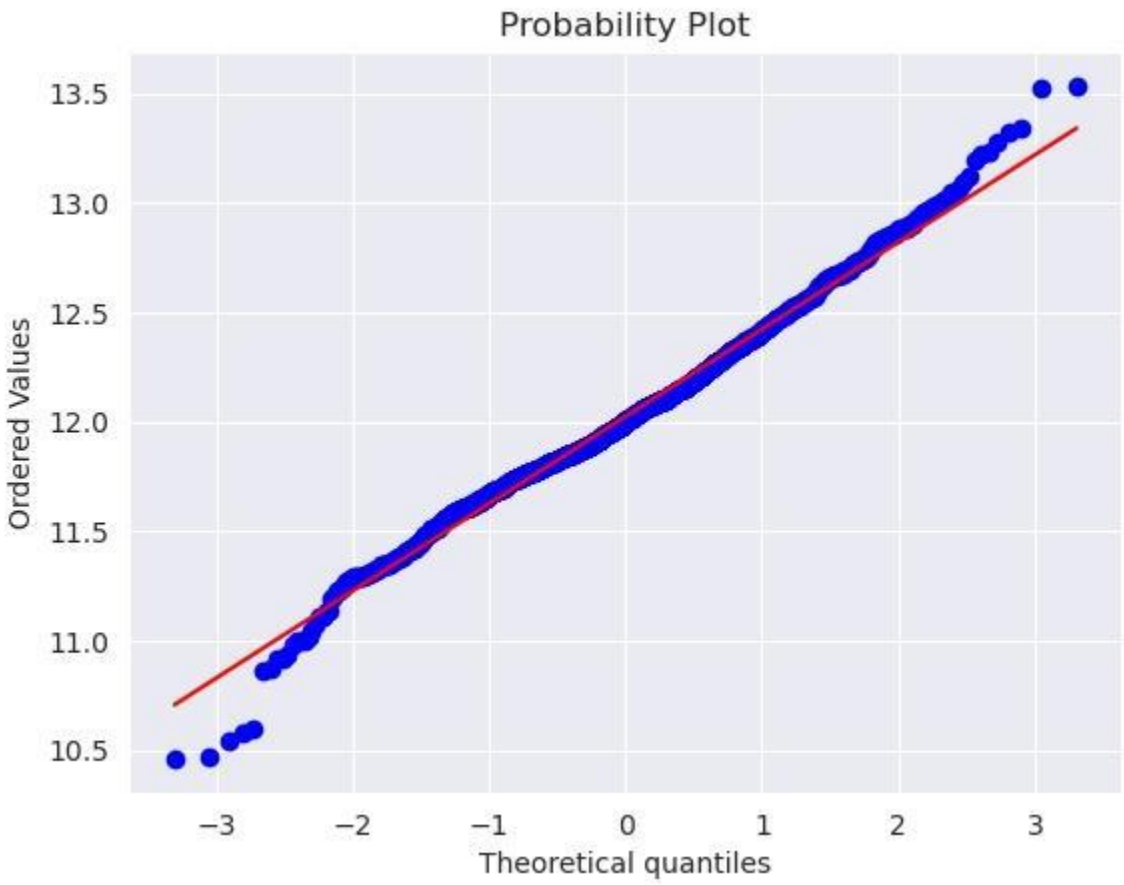

```
train["SalePrice"] = np.log1p(train["SalePrice"])

sns.distplot(train['SalePrice'] , fit=norm);
```

```
(mu, sigma) = norm.fit(train['SalePrice'])
print( '\n mu = {:.2f} and sigma = {:.2f}\n'.format(mu, sigma))

plt.legend(['Normal dist. ($\mu=$ {:.2f} and $\sigma=$ {:.2f} )'.format(mu, sig
ma)],
            loc='best')
plt.ylabel('Frequency')
plt.title('SalePrice distribution')


fig = plt.figure()
res = stats.probplot(train['SalePrice'], plot=plt)
plt.show()
```



The skew seems now corrected and the data appears more normally distributed…

CONCLUSION

Thus, the machine learning model using linear regression algorithm is very helpful in predicting the house prices for real estate customers