SINGAPORE FLAT RESALE PRICE PREDICTION WITH MACHINE LEARNING – RANDOM FOREST REGRESSOR

PROBLEM STATEMENT:

The objective of this project is to develop a machine learning model and deploy it as a user-friendly web application that predicts the resale prices of flats in Singapore. This predictive model will be based on historical data of resale flat transactions, and it aims to assist both potential buyers and sellers in estimating the resale value of a flat.

TOOLS USED:

- PYTHON - GOOGLE COLAB
- MACHINE LEARNING
- STREAMLIT
- GITHUB

IMPORT LIBARARIES:

Import libraries for handle the given data and preprocessing for the model training

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
import statistics
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
import pickle
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
```

DATA SOURCE:

We have a 4csv files each representing the specific time. The time periods are 1990 - 1999,2000-2012 ,2012-2014, 2015-2016, 2017 onwards. We want to setup these datasets as a one data frame

```
data =
pd.read_csv(r'C:\Users\LENOVO\Documents\dataset\ResaleFlatPricesBasedonAppr
ovalDate2000Feb2012 (1).csv')
data1=
pd.read_csv(r'C:\Users\LENOVO\Documents\dataset\ResaleFlatPricesBasedonAppr
ovalDate19901999 (2).csv')
data2 =
pd.read_csv(r'C:\Users\LENOVO\Documents\dataset\ResaleFlatPricesBasedonRegi
strationDateFromJan2015toDec2016 (1).csv')
data3=pd.read_csv(r'C:\Users\LENOVO\Documents\dataset\Resaleflatpricesbased
onregistrationdatefromJan2017onwards (1).csv')
data4=pd.read_csv(r'C:\Users\LENOVO\Documents\dataset\ResaleFlatPricesBased
onRegistrationDateFromMar2012toDec2014 (1).csv')
```

```
dataframe=(data,data1,data2,data3,data4)
dataset=pd.concat(dataframe)
```

Data source link: https://beta.data.gov.sg/collections/189/view

DATA PREPROCESSING:

After merge all these data we want to clean our data for better accuracy while predicting the resale price.

DROP NULL VALUES:

```
 dataset.isnull().sum()
dataset.dropna(inplace=True)
```

TYPECASTING THE NEEDED COLUMN:

```
dataset['flat_type']=dataset['flat_type'].astype(str)
dataset['flat_model']=dataset['flat_model'].astype(str)
dataset["resale_price"]=dataset['resale_price'].astype("float")
dataset['floor_area_sqm'] = dataset['floor_area_sqm'].astype('float')
dataset['lease_commence_date'] =
dataset['lease_commence_date'].astype('int')
dataset['lease_remaining_year'] = 99 - (2024 -
dataset['lease_commence_date'])
dataset['flat_type']=dataset['flat_type'].astype(str)
```

ENCODE THE COLUMN AS NUMERIC VALUE:

```
town = dataset['flat_model']
label_encoder = LabelEncoder()
dataset['flat_model']= label_encoder.fit_transform(town)

type_ = dataset['flat_type']
dataset['flat_type']= label_encoder.fit_transform(type_)
```

FROM THE STOREY_RANGE COLUMN FIND THE MEDIAN AND SAVE THE COLUMN AS STOREY_MEDIAN:

```
def get_median(x):
  split_list= x.split('TO')
  float_list= [float(i) for i in split_list]
  median= statistics.median(float_list)
  return median
dataset['storey_median']= dataset['storey_range'].apply(lambda
x:get_median(x))
```

CHECKING THE NULL VALUES IN THE COLUMN USING THE BOX PLOT:

```
column=['storey_median','lease_remaining_year','flat_model','flat_type','fl
oor_area_sqm',"resale_price"]
for i in column:
  plt.figure(figsize=(8,6))
  sns.boxplot(data=dataset,x=i)
  plt.title(f'box plot {i}')
  plt.xlabel(i)
  plt.show()
```

REMOVED THE OUTLIERS IN THE COLUMN USING THE QUANTILE METHOD:

```
def remove_outliers(df,column,multiplier=1.5):
  q1=df[column].quantile(0.25)
  q3=df[column].quantile(0.75)
  iqr=q3-q1
  lowwer_bound=q1-(iqr*multiplier)
  upper_bound=q3+(iqr*multiplier)
  df_cleaned=df[(df[column]>=lowwer_bound)&(df[column]<=upper_bound)]
  return df_cleaned
```

TRAIN TEST SPLIT:

```
X=df_cleaned[['floor_area_sqm','flat_type','flat_model','storey_median','le
ase_remaining_year']]
y =df_cleaned['resale_price']

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_sta
te=42)
```

INITIALIZE AND FIT THE RANDOM FOREST REGRESSOR:

To increase the accuracy of the prediction first normalise the data using RandomforestRegression() . Setup the new data as a trained 80% and tested data 20%. Here the dependent variable(y). While other features are independent variable

```
model=RandomForestRegressor()
model.fit(X_train,y_train)
```

PREDICTION USING TEHE REGRESSION MODEL AND SAVED THE TRAINED MODEL AS A PICKLE FILE:

```
prediction=model.predict(X_test)
with open('Random_Forest_model.pkl', 'wb') as model_file:
    pickle.dump(model, model_file)
```

MODEL EVALUATION:

Now it's time to check the Accuracy of our machine learning model .I have check various evaluation metrics those are mean_squared_error(MSE) - 7185531065.66,

mean_absolute_error(MAE)- 62366.18, r2_squared score – 0.74 it means the accuracy of the model is 74%

```
mse =mean_squared_error(y_test,prediction)
mae=mean_absolute_error(y_test,prediction)
r2=r2_score(y_test,prediction)
print('mean_squared_error:',mse)
print('mean_absolute_error:',mae)
print('r2_score:',r2)
```

CONCLUTION:

In conclusion, predicting flat prices is a complex task that involves analyzing numerous variables such as flat type, flat model, storey range , lease commence date, floor area sqm and resale price. While various methods, including statistical models, machine learning algorithms, and expert analysis, can be employed for price prediction.