Predicting Housing Purchase Decisions With Machine Learning

WONG WING YIN RENEE

Overview

Objective

Develop predictive models to forecast whether a customer is likely to purchase a property

- Helping real estate agents to recommend the most suitable properties to clients
- increase clients‘ purchase probability
- Reduce unnecessary property visits and improve agents' work efficiency

Dataset Composition & Feature Landscape

Global_House _Purchase

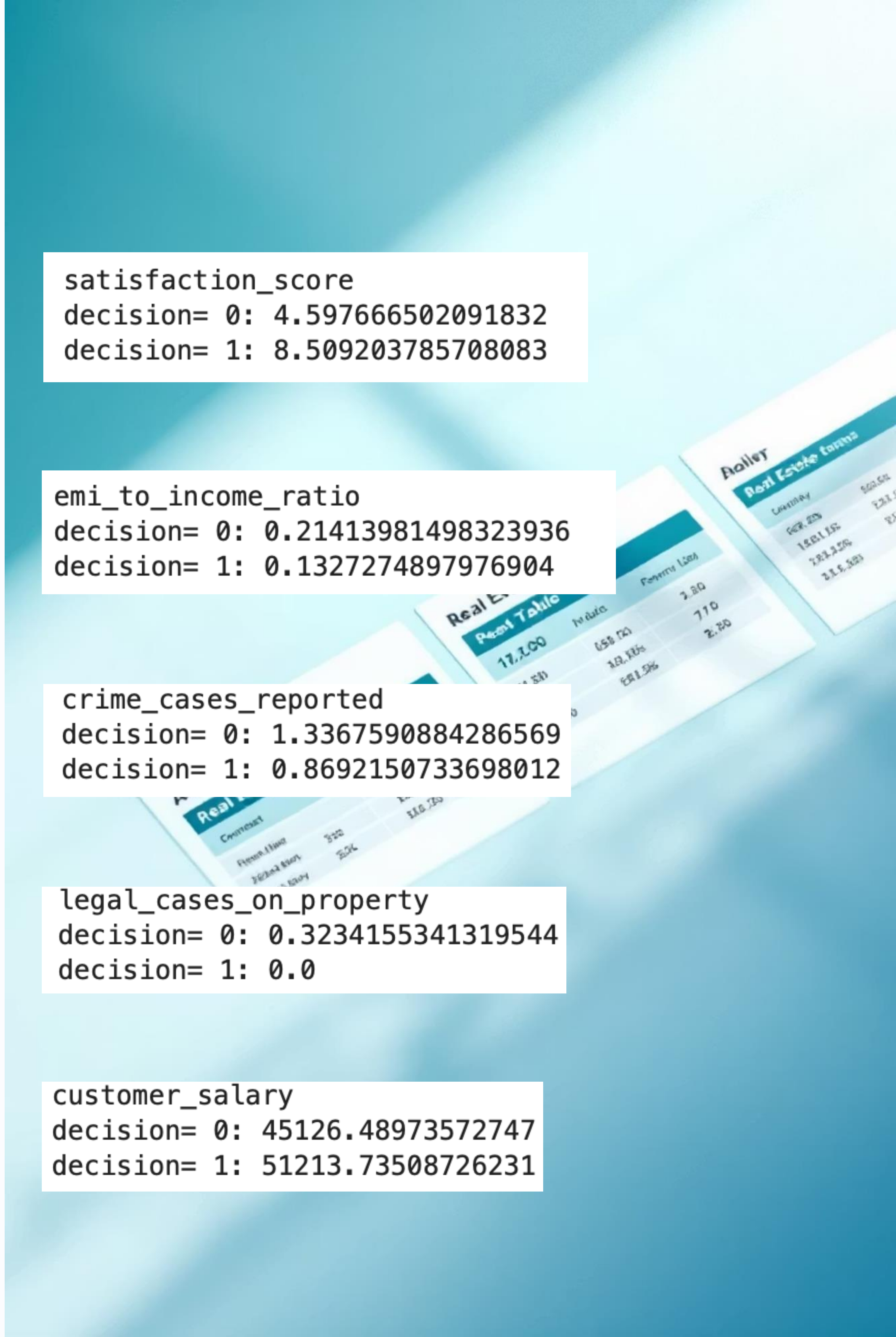
200000 records with 25 feature (demographics, property details)
Target variable : ‘decision’

Region Index

Country-level indicators such as GRY, affordability index, and price-to-income ratios

Key Features

- satisfaction_score
- emi_to_income_ratio
- crime_cases_reported
- legal_cases_on_property
- customer_salary



```
satisfaction_score
decision= 0: 4.597666502091832
decision= 1: 8.509203785708083
```

```
emi_to_income_ratio
decision= 0: 0.21413981498323936
decision= 1: 0.1327274897976904
```

```
crime_cases_reported
decision= 0: 1.3367590884286569
decision= 1: 0.8692150733698012
```

```
legal_cases_on_property
decision= 0: 0.3234155341319544
decision= 1: 0.0
```

```
customer_salary
decision= 0: 45126.48973572747
decision= 1: 51213.73508726231
```

Data Preprocessing

Dataset Merging

Combined Global_House _Purchase and Region Index CSV using country as the join key

```
df = pd.merge(df1, df2, on='country', how='inner')
```

Dropping unnecessary column

Dropping unnecessary columns to enhance model speed and reduce overfitting

```
df = df.drop(columns=['property_id','country','city','Rank'])
```

One-Hot Encoding for Categorical Variable

transform categorical features into numeric variables, allowing the model to interpret

```
df = pd.get_dummies(df, columns=['property_type',
                                 'furnishing_status'], drop_first=True)
```

Data Cleaning

Fill in all the missing values

Heatmap and logistic regression analysis: Correlation Insights

Heatmap

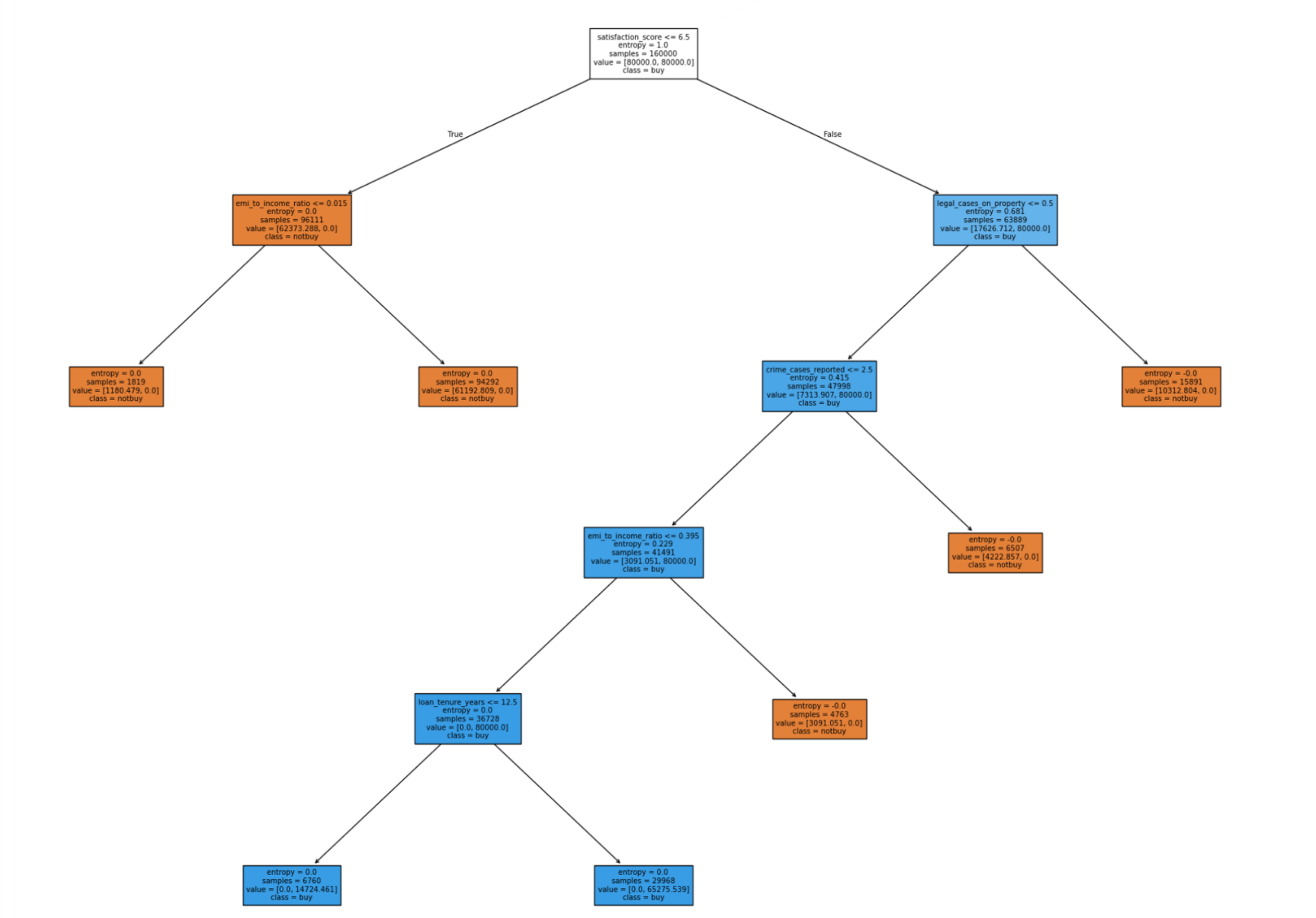
```
with「decision」top 5 positive correlation :
satisfaction_score                            0.572783
customer_salary                               0.091546
Price to Rent Ratio City Centre               0.039186
Price to Rent Ratio Outside Of City Centre    0.038722
loan_tenure_years                             0.022259
Name: decision, dtype: float64

with「decision」top 5 negative correlation :
legal_cases_on_property                -0.314936
crime_cases_reported                   -0.166080
emi_to_income_ratio                    -0.156033
property_size_sqft                     -0.057232
Gross Rental Yield Outside Of Centre   -0.049401
```

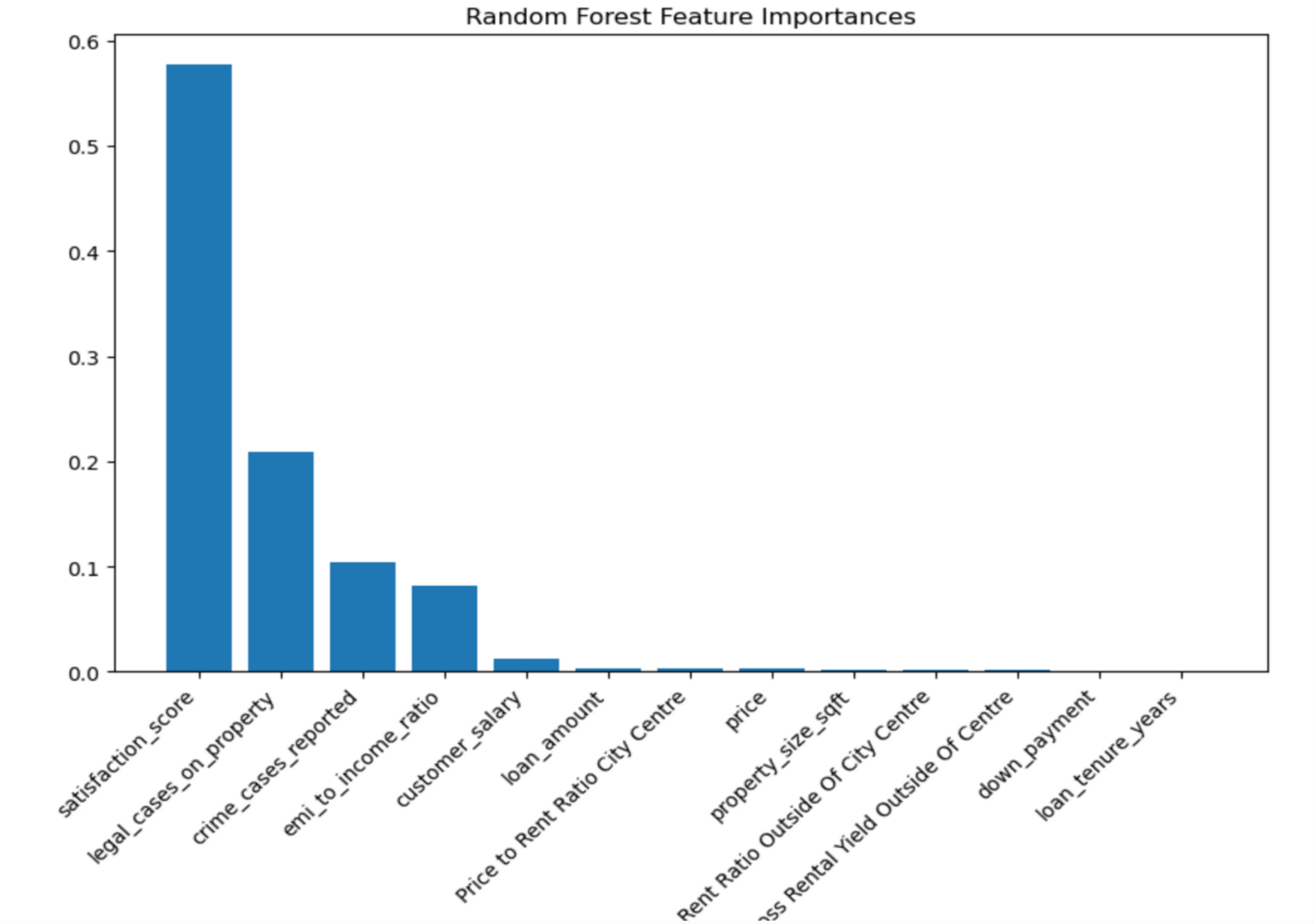
Logistic Regression

	Feature	Coefficient
16	satisfaction_score	17.200346
11	loan_amount	2.038655
1	price	1.272431
0	property_size_sqft	0.901294
23	Price to Rent Ratio Outside Of City Centre	0.525942

	Feature	Coefficient
14	down_payment	-0.806004
10	customer_salary	-1.943466
8	crime_cases_reported	-14.960176
9	legal_cases_on_property	-16.908158
15	emi_to_income_ratio	-47.825768

Tree-Based Models: Perfect Performance





1.00

Accuracy

Decision Tree & Random Forest

1.00

ROC AUC

Perfect discrimination

57.65%

Feature Importance

satisfaction_score dominated

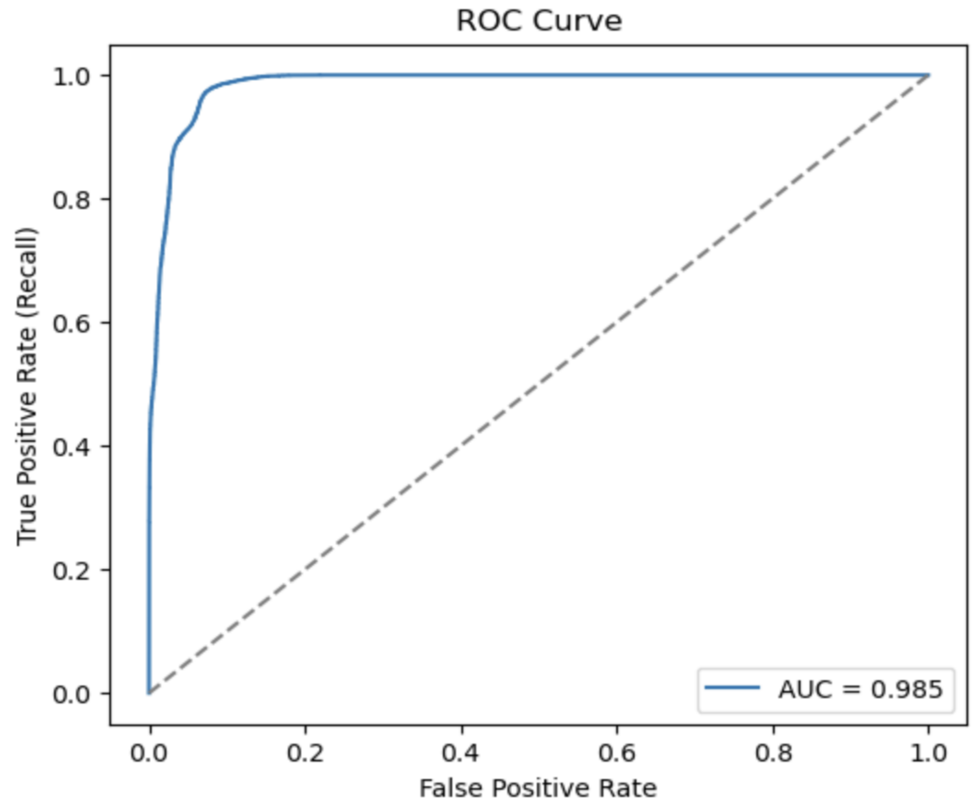
Initial Logistic Regression Modeling

Logistic Regression Results

Models with all features achieved unrealistically high performance due to the feature “satisfaction_score”, which is only available after property viewing.

```
Train Accuracy is: 93.82%
Test Accuracy is: 93.84%
```

```
Confusion Matrix for logisticRegression:
=================
TN= 28548 FP= 2112 FN= 325 TP= 9015
Recall=  0.965
Specificity=  0.931
Precision=  0.81
F1-score: 0.8809302780085015
```

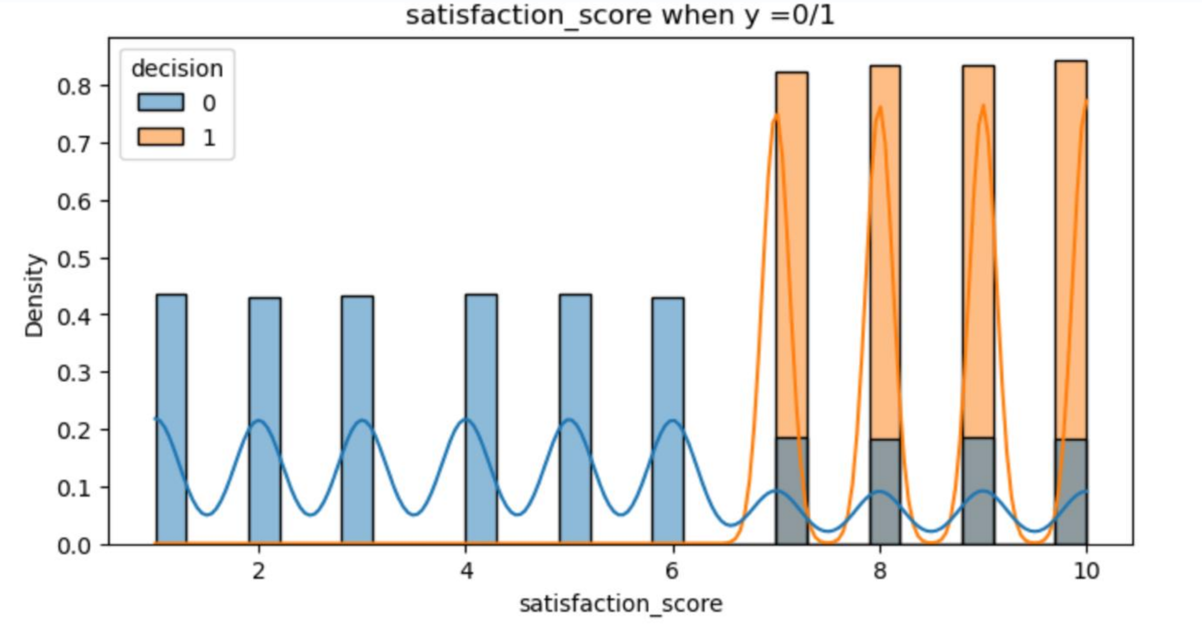



```
ROC AUC Score: 0.9852526990086756
```

Solution: Remove Features and Class Balancing

Remove Subjective Features

Excluded satisfaction_score (for logistic regression, MLP), and legal_cases_on_property, crime_cases_reported, and emi_to_income_ratio (for decision tree and random forest models) in the following models, making them more closely reflect real-world scenarios and preventing the models from being dominated by a single or small set of features.




```
decision
0    153932
1     46068
Name: count, dtype: int64
```

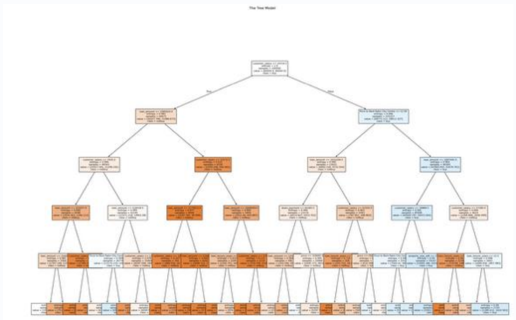
Apply SMOTE

Generating minority class samples to balance the class distribution and improve predictions for the minority class.

MLP result

```
Accuracy: 0.746875
ROC AUC: 0.7736503804243824
              precision    recall  f1-score   support

           0       0.78      0.92      0.85     30660
           1       0.40      0.17      0.24      9340

    accuracy                           0.75     40000
   macro avg       0.59      0.55      0.54     40000
weighted avg       0.69      0.75      0.71     40000
```

Model Comparison & Strategic Recommendations



Model	Accuracy	Recall (1)	F1-Score	ROC AUC	Status
Logistic Regression	0.67	0.90	0.56	0.77	Moderate
Random Forest	0.49	0.75	0.4	0.60	poor generalization
MLP	0.69	0.65	0.50	0.77	Moderate

1 Application value

Logistic Regression achieves a higher recall and F1-score for class 1, making it better and faster for this case

2 Data Collection

Focus on pre-viewing variables with strong correlation: crime statistics, income ratios

3 Limit dependency on satisfaction score

Satisfaction score is a post-event feature and should not be heavily relied upon for initial predictions. It is more suitable for refining subsequent property recommendations.

4 Continuous Improvement

Incorporate geographic features (nearby transport, mall, and recreation facility)

Thank you