UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

MACHINE LEARNING
Regression & Classification

31st May 2025

PHUONG CHI DAO (474064)
I PUTU AGASTYA HARTA PRATAMA (472876)

Classification: Insurance Claims

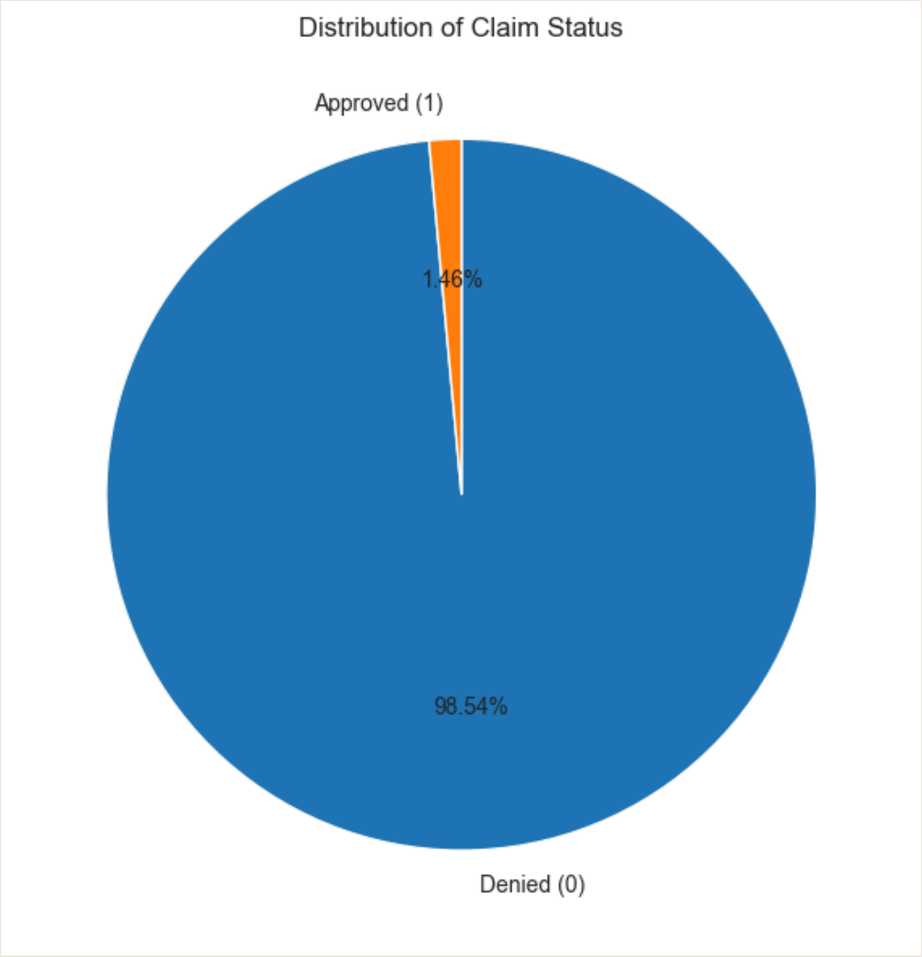
A dataset comprises **56,993 observations,** with the dependent variable being the claim status, which is categorized as 0 (denied) and 1 (approved).

Out of all the claims received, **only 1.46% were approved**, which reflects the reality of the business case. However, this makes the data extremely imbalanced, requiring careful consideration.

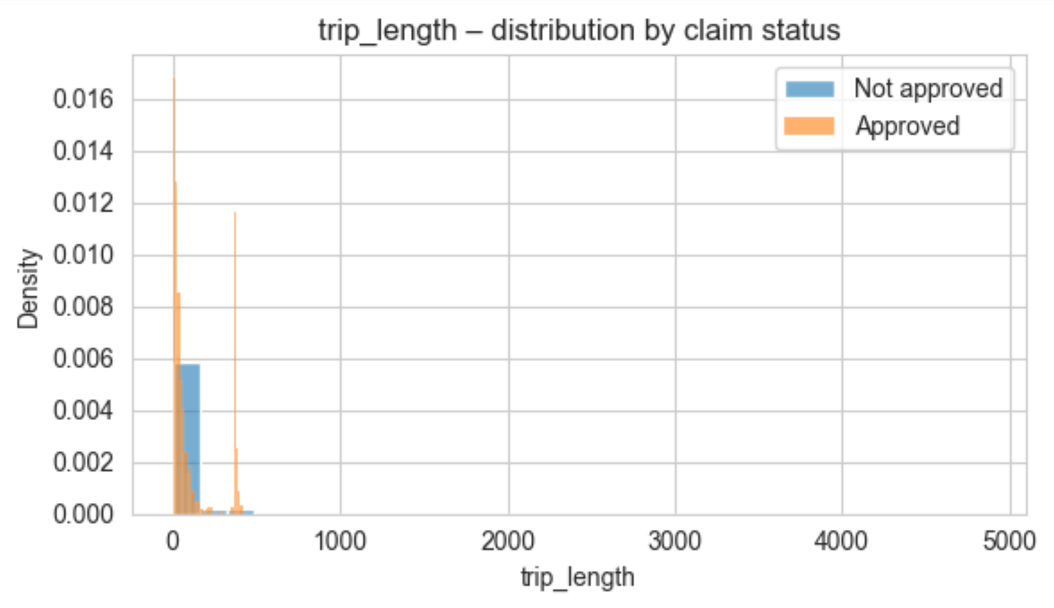
Key features:

Numeric: reward, revenue, person_age, customer_score, trip_length, support_interactions

Categorical: product_id, entity_type, channel, agent_id, location, person_gender

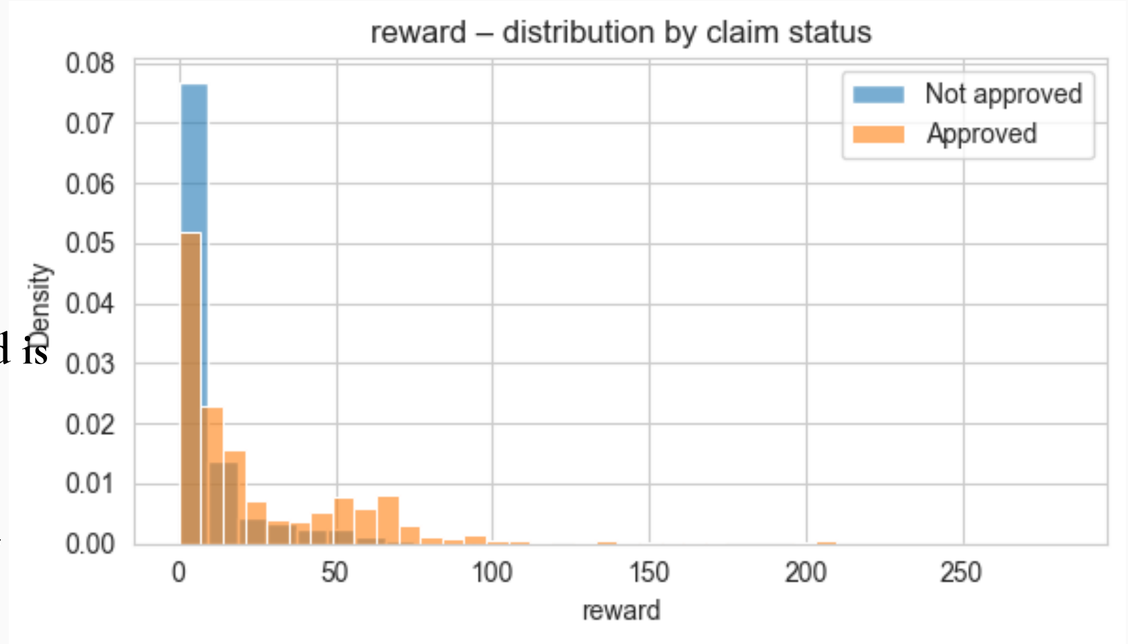


Insurance Claims: Data Overview

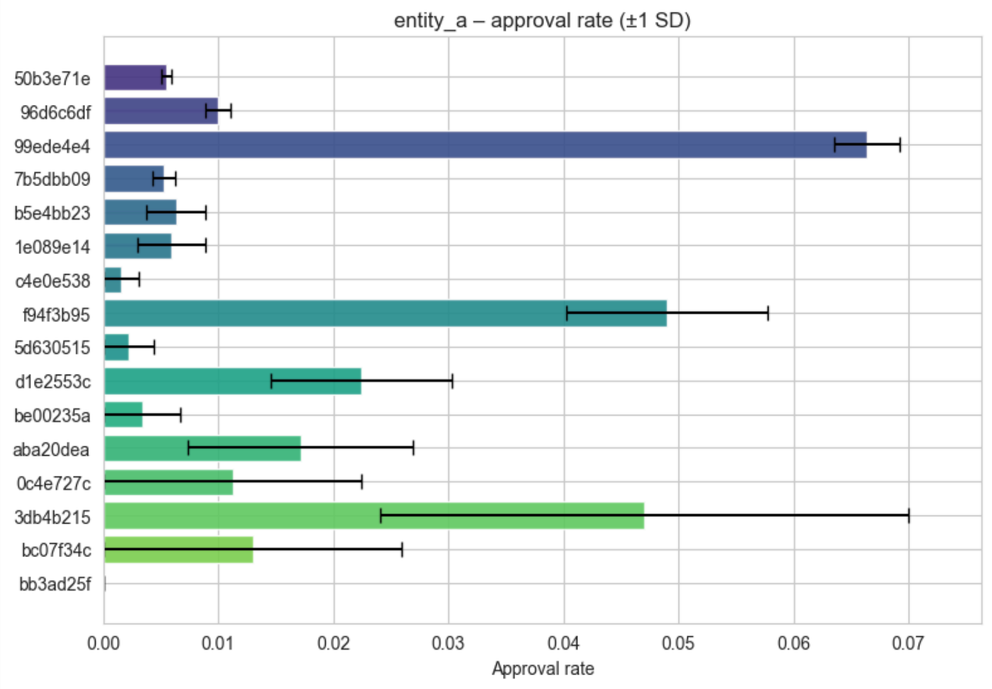


Length of trip is right-skewed (max 4856 days vs 75th percentile: 53 days)

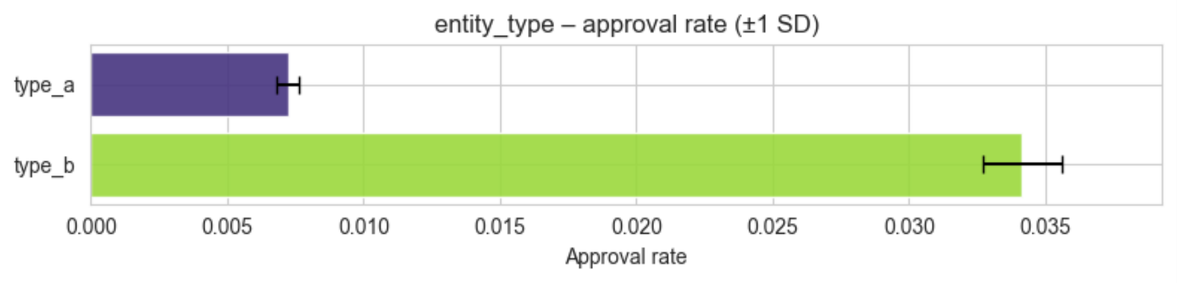
The probability of a claim with a higher reward being accepted is much higher than that of claims with lower rewards



Categorical features by claim status

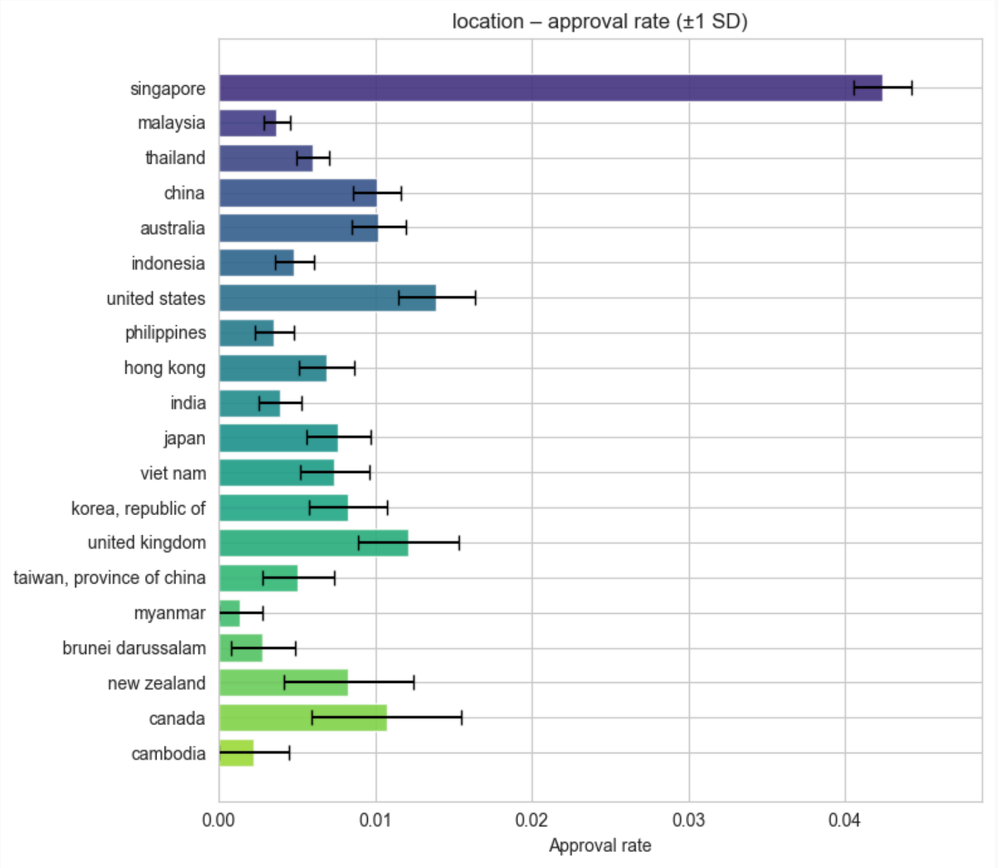


Three dominant entities received higher claim approval
99ede4e4 - f94f3b95 - 3db4b215

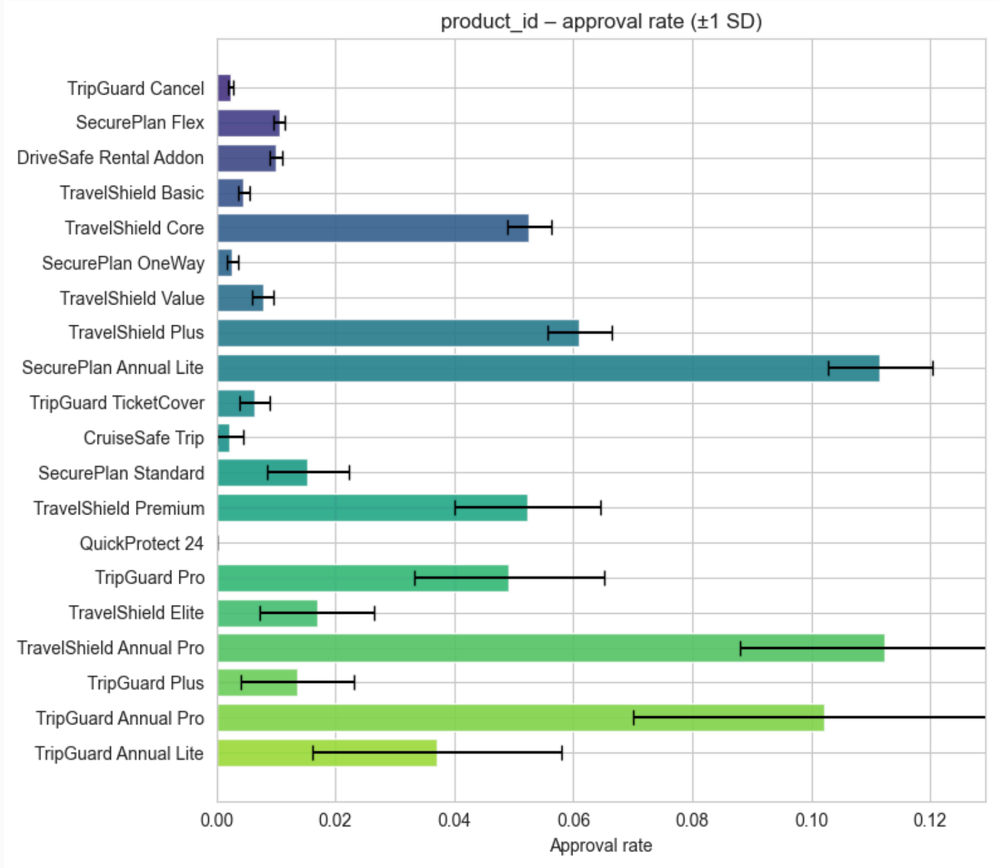


Entity type B has a significantly higher approval claim rate

Categorical features by claim status



The majority of approval cases are based in Singapore, which could indicate that the insurance company is located in Singapore or that the majority of their clients are from Singapore. Followed by the United States, Canada, and the United Kingdom.



Annual insurance product received the highest claim approval, including TravelShield Annual Pro, SecurePlan Annual Lite, Tripguard Annual Pro

Key Findings

Annual Policies

Represent 2.3 % of all insurances (1,309 out of 56,993)

Predominantly sold in Singapore (1,115 cases, 85 % of annuals)

Agent Concentration

Three agents handle nearly all long-haul plans (agent 5, 13, 14)

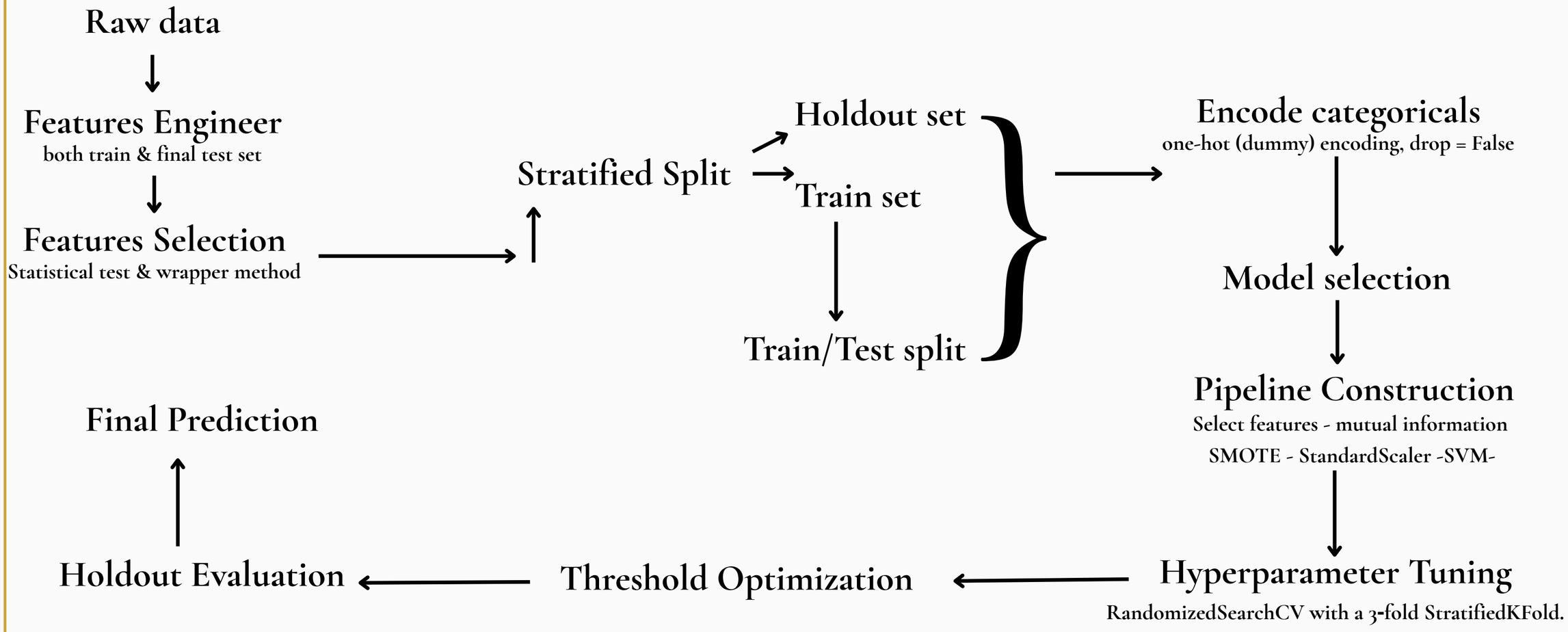
Entity

Type B has more positive approves than type A

Product-Level Approval Rates

The annual insurance product received the highest claim approval, with Travel Shield Annual Pro at an 11.7% approval rate

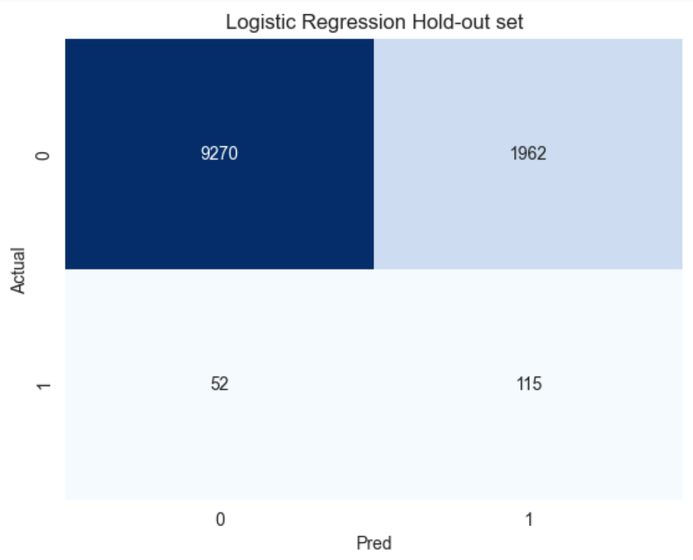
Modelling Approach

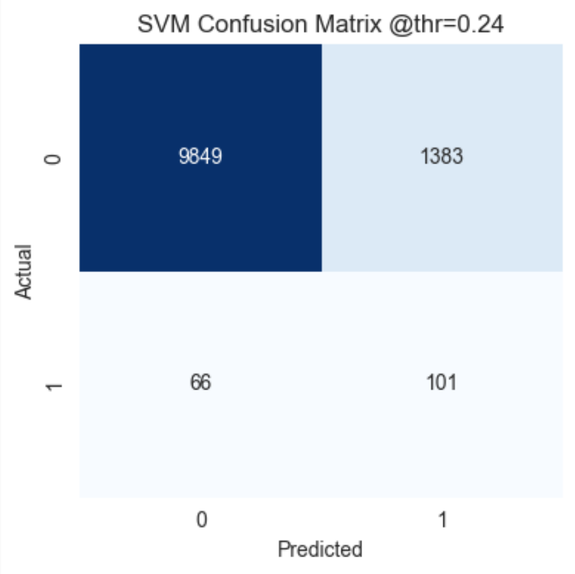


Model comparison

	balanced_acc	precision	recall	f1
svm_model	0.757	0.063	0.659	0.115
lr_model	0.701	0.084	0.479	0.143
knn_model	0.644	0.041	0.443	0.075
xgboost_model	0.593	0.062	0.240	0.099
rf_model	0.527	0.054	0.072	0.062

Table 1: Model comparison





Why SVM?

Final result with hold-out set

```
Hold-out metrics @threshold=0.24
Balanced-Accuracy: 0.741
Precision : 0.068
Recall: 0.605
F1 Score: 0.122
```

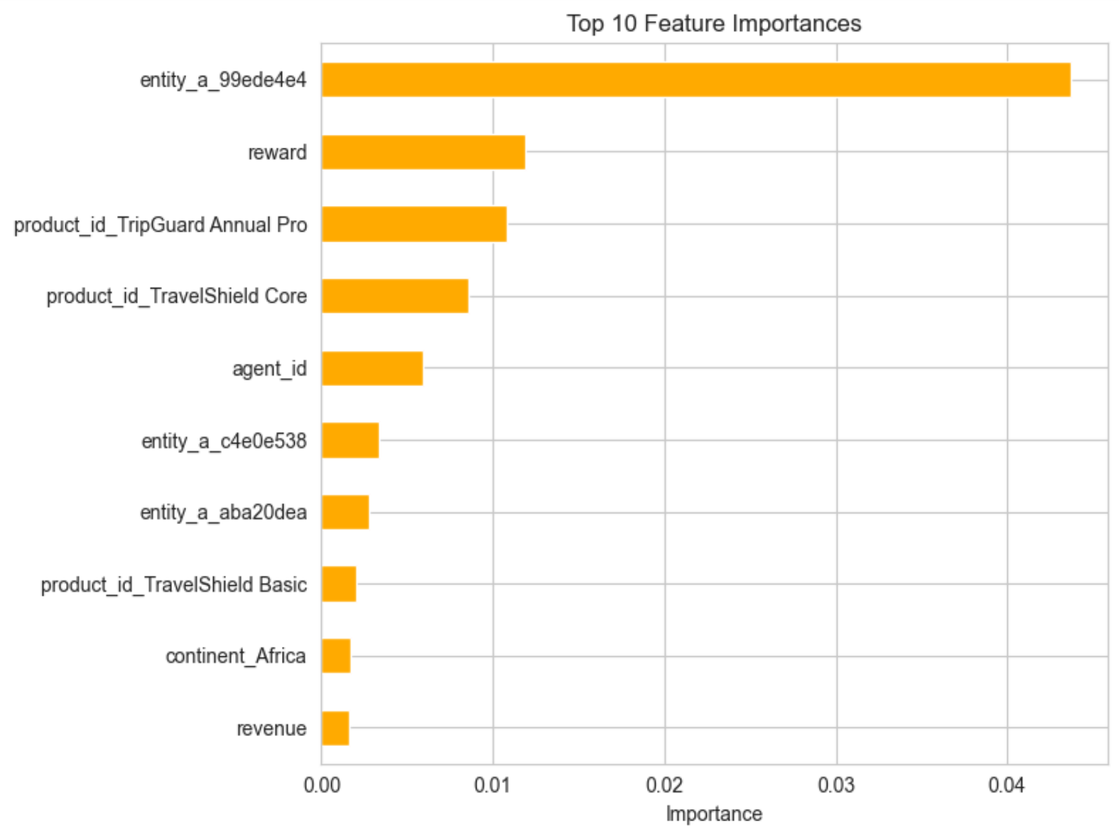
- Best recall for the minority class
- Effective in high-dimensional space
- Robust to overfitting with RBF kernel

Final insurance claim prediction

5513 - non-approved claim status

820 - approved claim status

Features important



- Entity 99ede4e4 category drives the most variance in claim outcomes, echoing our earlier observation that certain underwriting entities (like those in Singapore) dominate approval rates
- Reward ranks second, confirming that higher-value incentives correlate strongly with claim approvals.
- Among product IDs, TripGuard Annual Pro and TravelShield Core appear in the top four
- Agent id is also highly important, reinforcing that a handful of agents (e.g., Agent 5, 13, 14) handle most high-approval policies

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

Regression: Prediction of Apartments Price

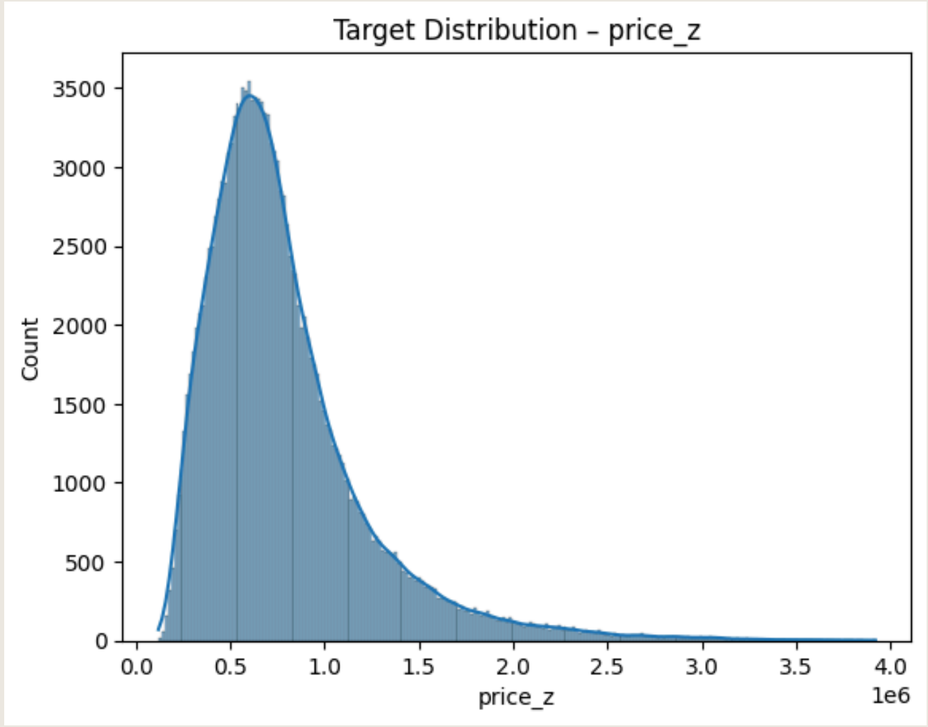
A dataset comprises **156,454 observations,** with the dependent variable being the apartment price (target_z).

Right-skewed target, price of apartment, according to certain attributes reflective of most of the housing are relatively priced on the low-end, while a small number of luxury or very well-located apartments push the average price higher.

Selected Key Features:

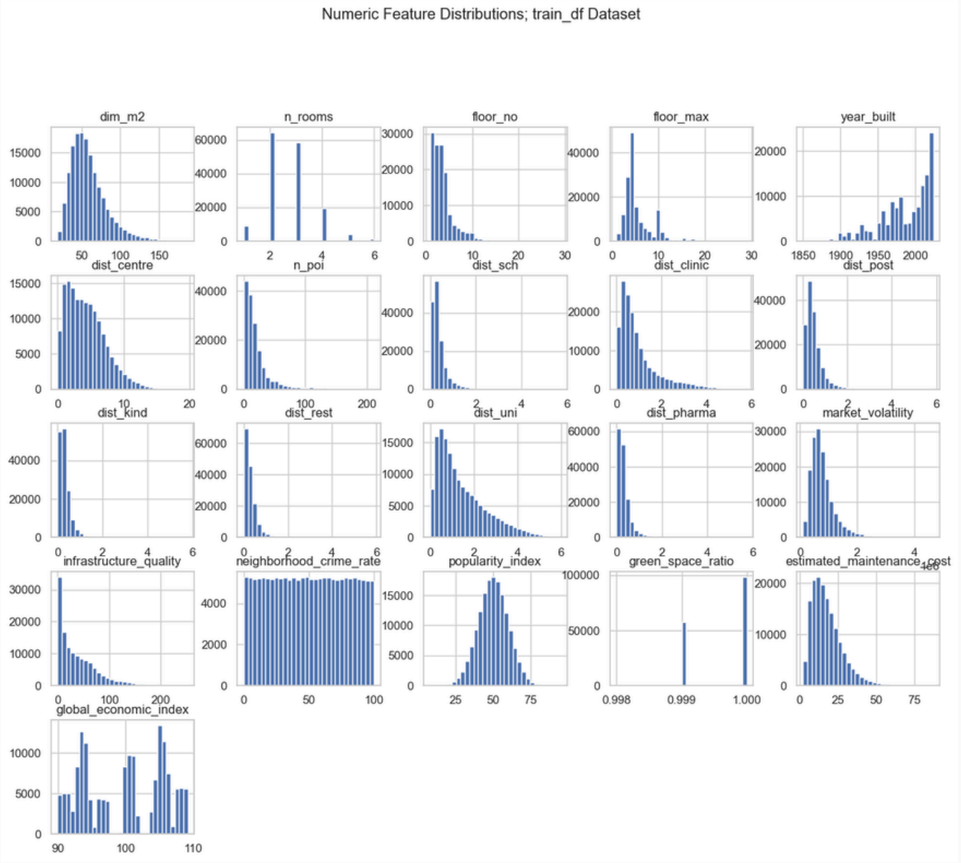
Numeric: 'market_volatility' 'dim_m2' 'estimated_maintenance_cost' 'n_poi', 'infrastructure_quality' 'dist_centre' 'dist_rest' 'dist_clinic', 'floor_max' 'dist_sch' 'dist_uni' 'year_built' (binned)

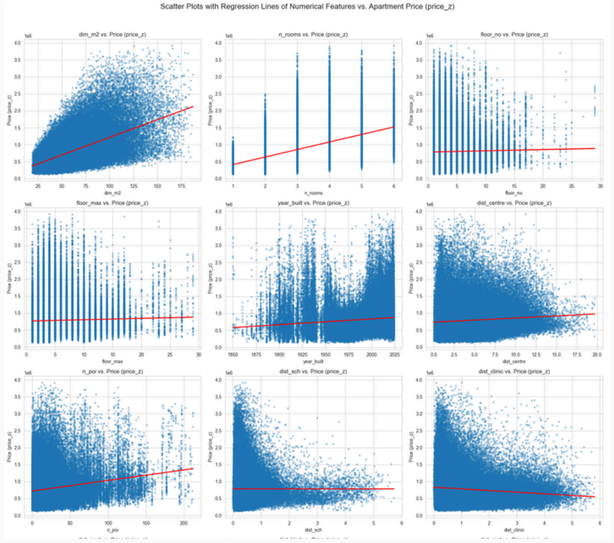
Categorical: 'obj_type', 'n_rooms', 'own_type', 'build_mat', 'has_park', 'has_balcony', 'has_lift', 'has_sec', 'has_store', 'loc_code', 'green_space_ratio' 'src_month' (binned)

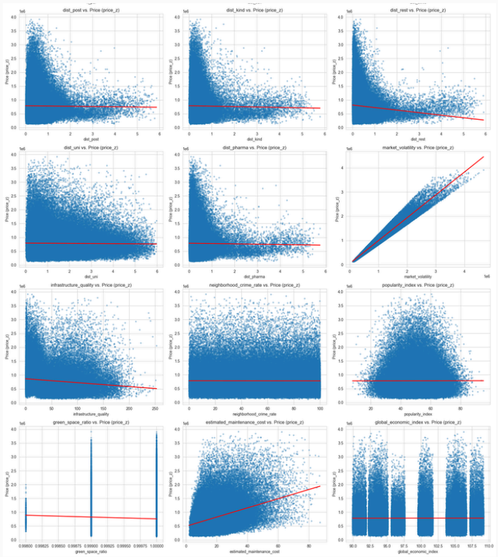


Regression: Prediction of Apartments Price

Analysis of Numerical Variables





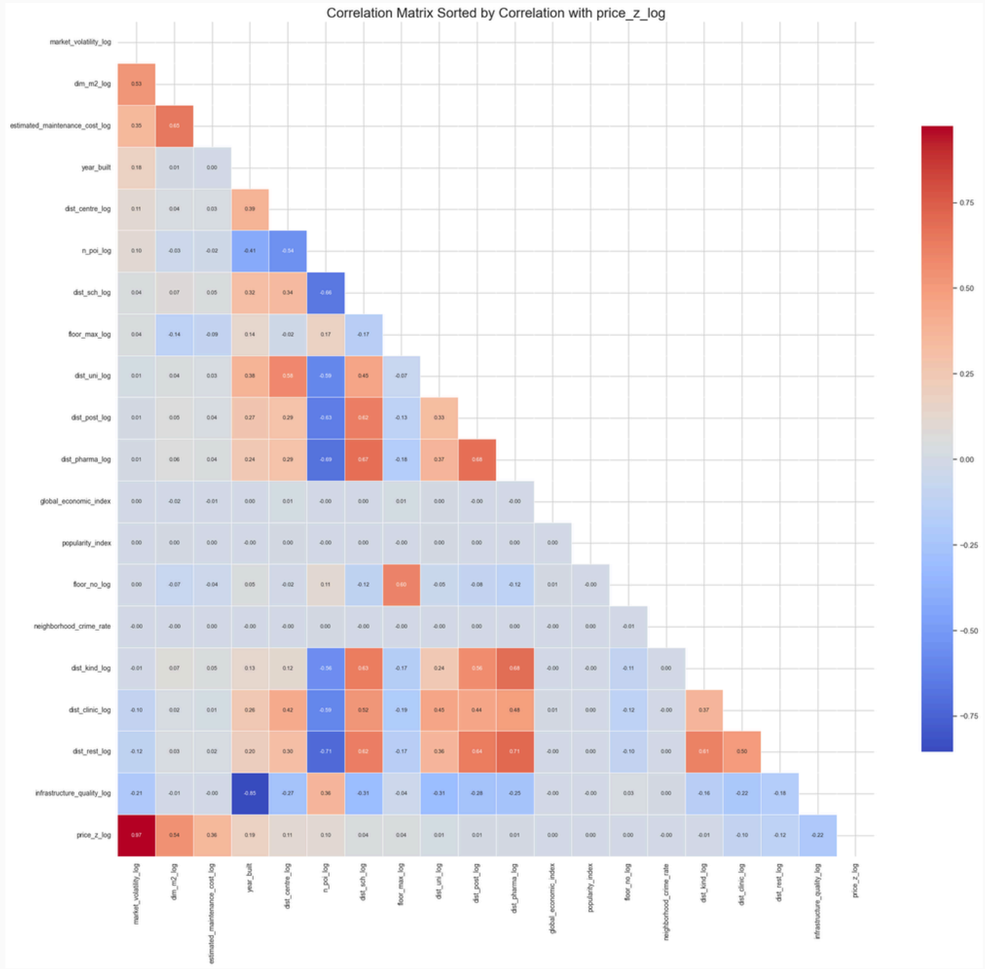


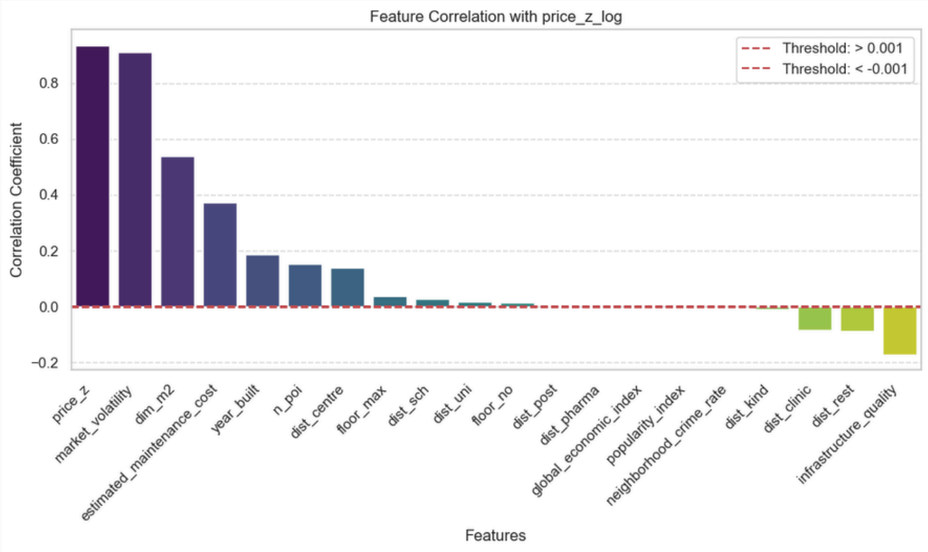
Most of the variables (including price) are right skewed, indicating that log transformation needed

Only a few variables have positive correlation with target → limited predictive power to the target

Regression: Prediction of Apartments Price

Correlation of Numerical Variables



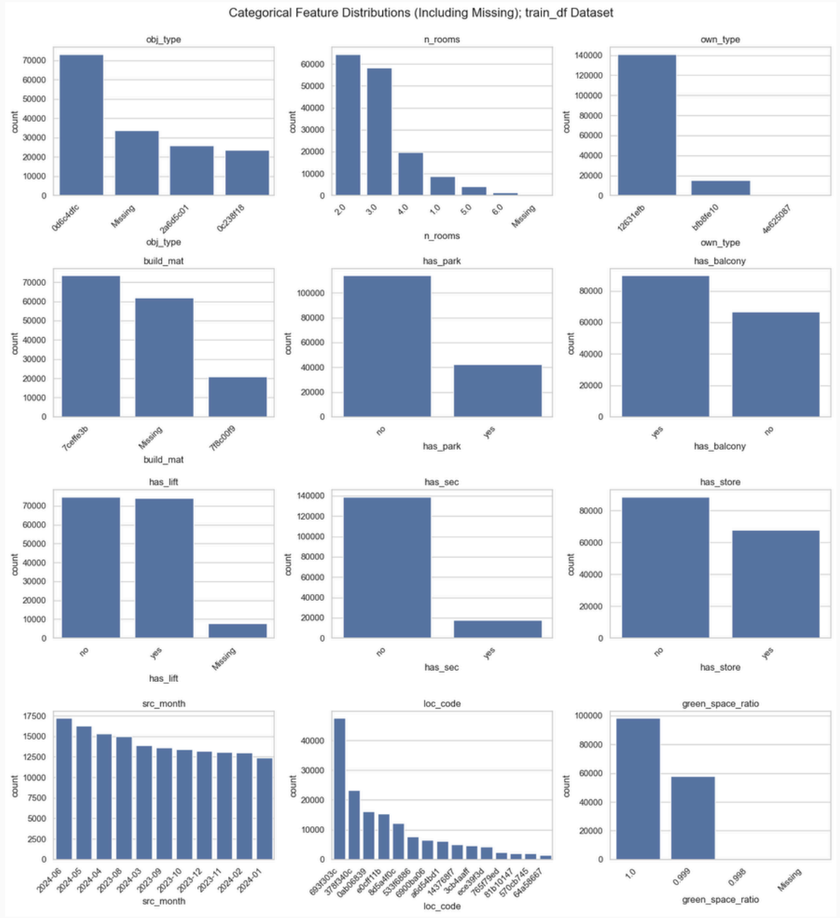


Numerical variables are then selected based on their pearson correlation in target

Threshold of 0.01 to avoid

Regression: Prediction of Apartments Price

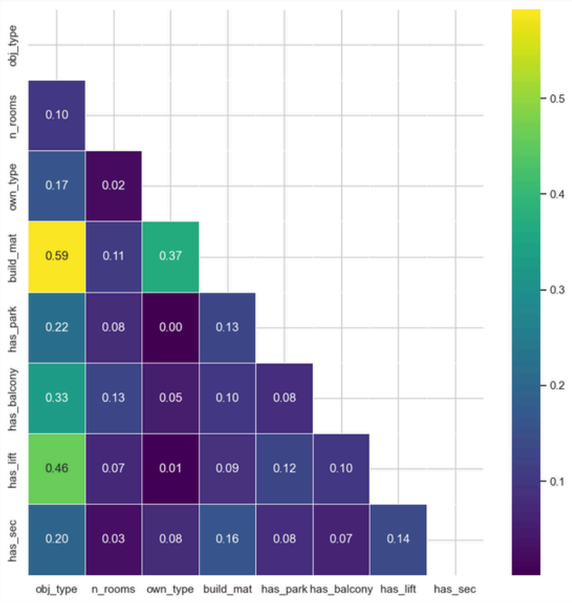
Analysis of Categorical Variables



ANOVA; Features & Target

	Variable	F-statistic	p-value
1	n_rooms	9153.386398	0.000000e+00
3	build_mat	7937.029873	0.000000e+00
0	obj_type	7481.079584	0.000000e+00
6	has_lift	5898.004374	0.000000e+00
8	has_store	4151.659578	0.000000e+00
10	loc_code	4016.960398	0.000000e+00
4	has_park	2928.606959	0.000000e+00
7	has_sec	2867.340981	0.000000e+00
2	own_type	890.181358	0.000000e+00
5	has_balcony	762.372912	2.071141e-167
11	green_space_ratio	442.085900	3.507697e-192
9	src_month	235.027819	0.000000e+00

Crammer's V: Multicolinearity of Features



No multicolinearity detected
All of the categorical features are statistically significant

Regression: Prediction of Apartments Price

Data Preprocessing: Treatment of Variables during Data Transformation

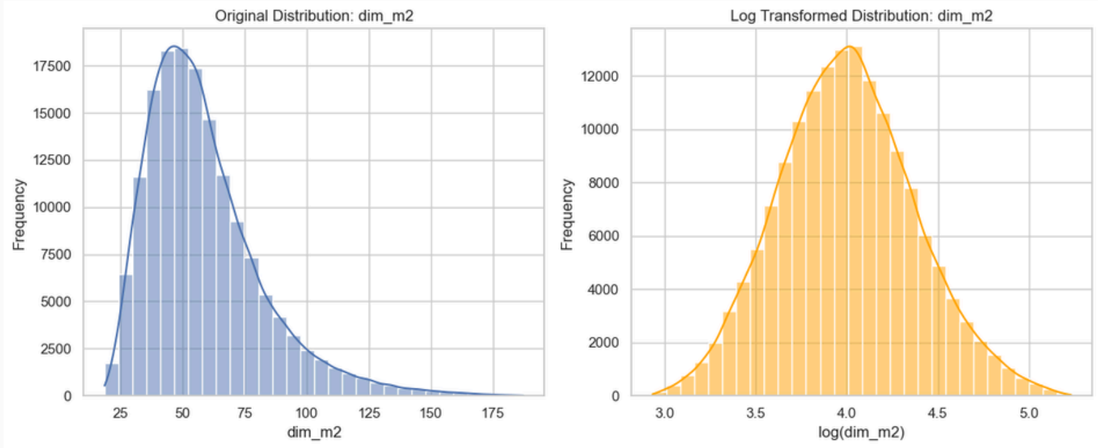
Imputation on Missing Value ONLY BASED on Train Set Then applied on Validation and Test set

- For numerical; Median Imputation
- For Categorical: Mode Imputation

Categorical: OneHotEncoding

has_sec_no	has_sec_yes	has_store_no	has_store_yes	loc_code_0ab06839	loc_code_143768f7	loc_code_378f340c
0.0	1.0	1.0	0.0	0.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	1.0	0.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	1.0	1.0	0.0	0.0

Numerical: Log Transformation to Normalise Distribution



Numerical: Scaling

- For OLS, Ridge, Lasso, Elastic Net: StandardScaler()
- For KNN: MinMaxScaler()
- For XGBoost: No need :)

Regression: Prediction of Apartments Price

Data Preprocessing: Features Engineering

Binning of Year and Month

```python
for df in [train_df, test_df]:
    df['year_built_bin'] = pd.cut(df['year_built'], bins=[1900, 1950, 2000, 2025],
                                  labels=["pre-1950", "1950-2000", "post-2000"])
    df['year_built_bin'] = df['year_built_bin'].astype('category')
    df['year_built_bin'] = df['year_built_bin'].cat.add_categories('Missing').fillna('Missi

    df['src_month_num'] = pd.to_datetime(df['src_month'], format='%Y-%m').dt.month
    df['src_month_bin'] = pd.cut(df['src_month_num'], bins=[0, 3, 6, 9, 12],
                                 labels=["Q1", "Q2", "Q3", "Q4"], include_lowest=True)
    df['src_month_bin'] = df['src_month_bin'].astype('category')

categorical_features.extend(['year_built_bin', 'src_month_bin'])
```

Python

Year:

- pre-1950 (old)
- 1950-2000 (somehow new)
- post-2000 (new building)

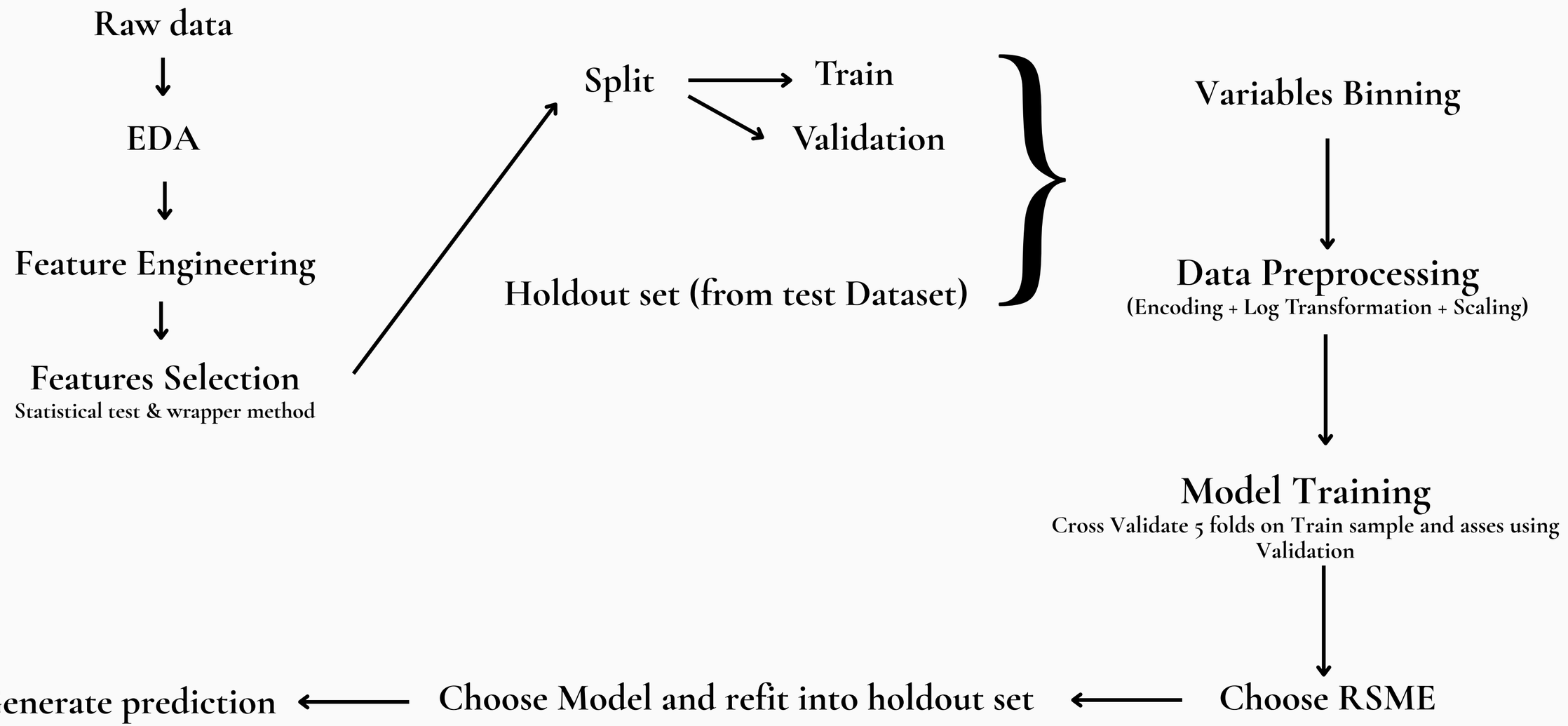
Month:

- Q1
- Q2
- Q3
- Q4

Binned quarterly to avoid seasonality

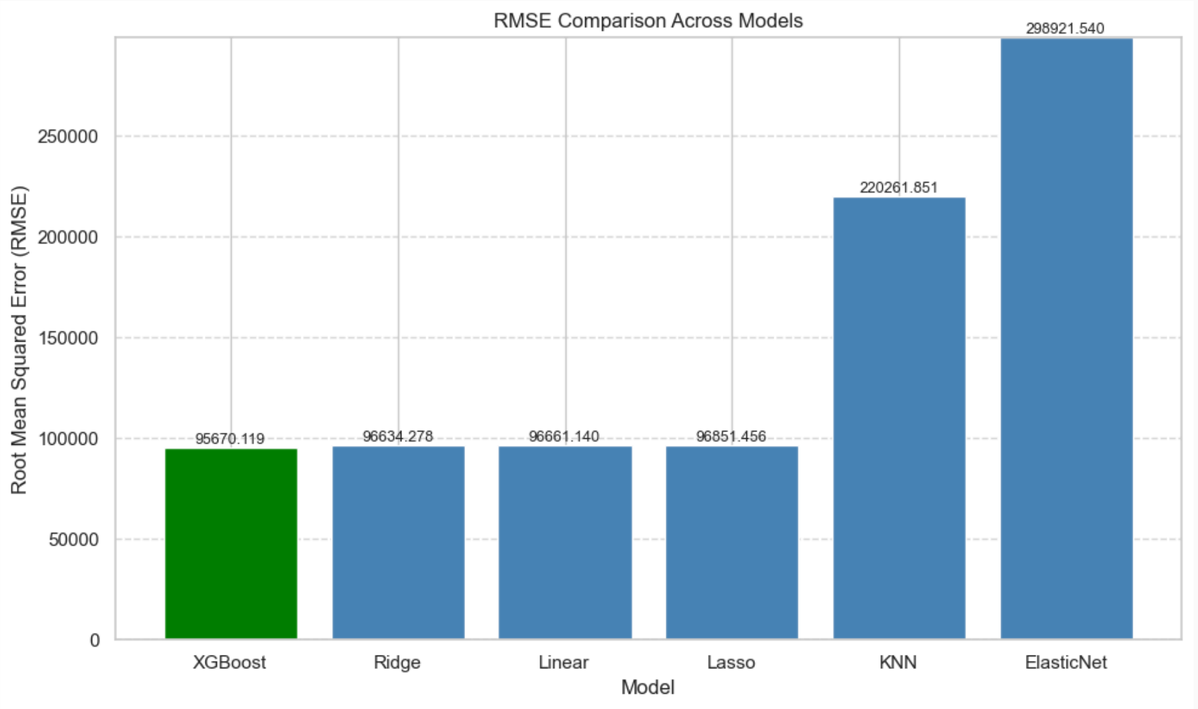
UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

Building Regression ML Pipeline



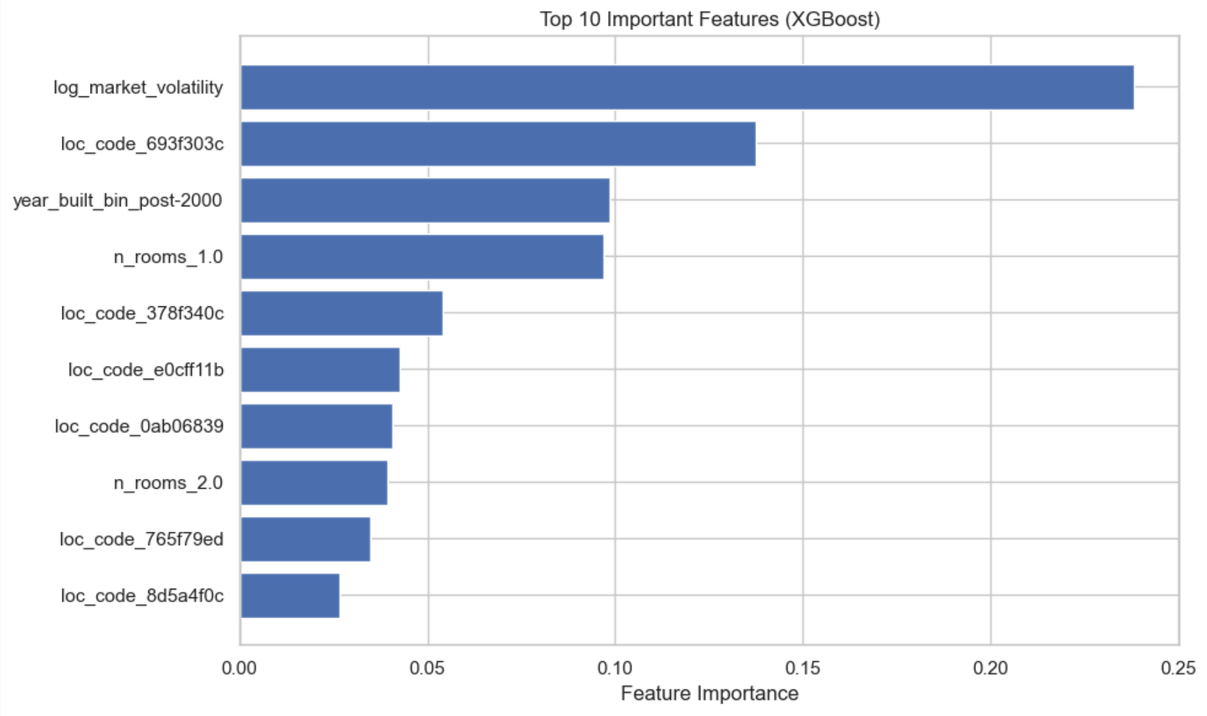
Model Selection

Model	RMSE	MAE	R2
Linear	96661.140	69894.471	0.951
Ridge	96634.278	69880.721	0.951
Lasso	96851.456	70230.531	0.951
ElasticNet	298921.540	186211.389	0.530
KNN	220261.851	146305.666	0.745
XGBoost	95670.119	68788.042	0.952



XGBoost achieved the lowest validation RMSE among all tested model

Features Importance



- Apartment prices are most strongly influenced by fluctuations or uncertainty in the market
- Geographic location is a strong predictor of apartment prices.
- Newer buildings seem to carry pricing weight
- Sum of room drives the prices up (assuming 1 room means studio)

Prediction on apartment price heavily influenced by macroeconomic, locational factors, and consumer preference dominate valuation signals

THANK YOU !!!