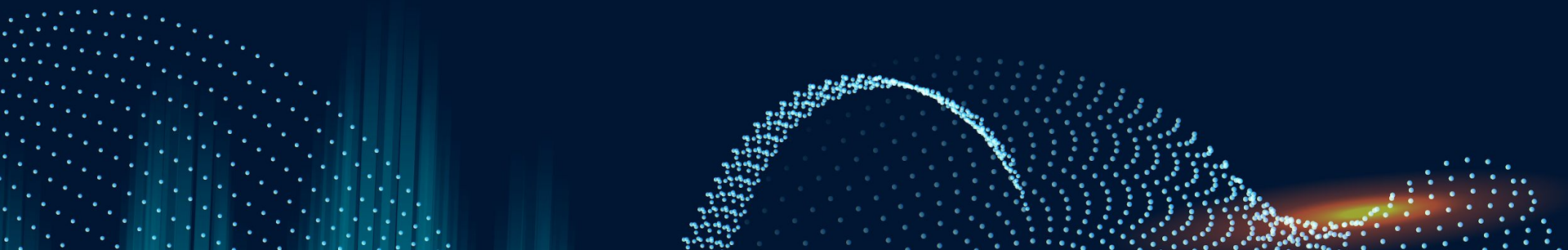
Automated Property Price Evaluation

Real estate data from **Restb.ai**

Andreja Andrejic
Jinheng Lin
Simone Paloschi
Alexandros Tremopoulos

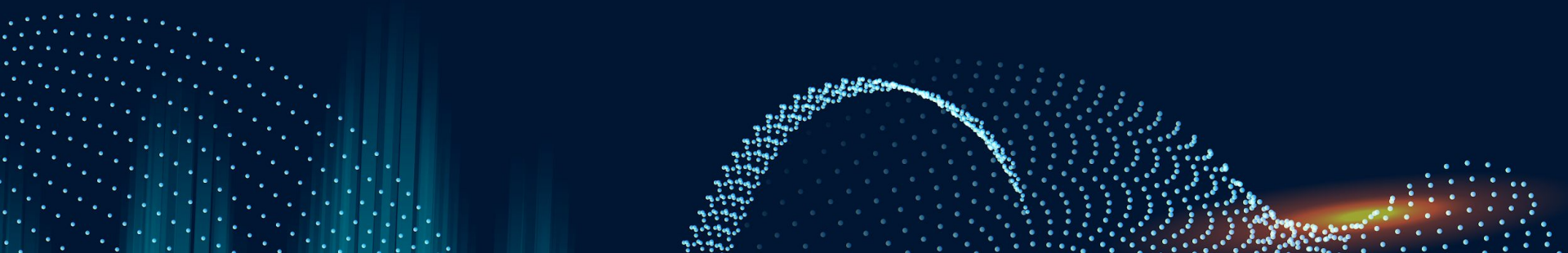
Dataset description

- **54** input variables - 1 target
- **631** variables after expansion
- **1.430.961** missing values (24% of total data)
- **107.437** rows of train
- **22.039** rows of test



Preprocessing

- Drop Columns → more than 85% missings
- Remove Rows → more than 20% missing
- Expansion of list variables → new binary column for each value from the list variable present in the training set (around 500 feats)



Formatting

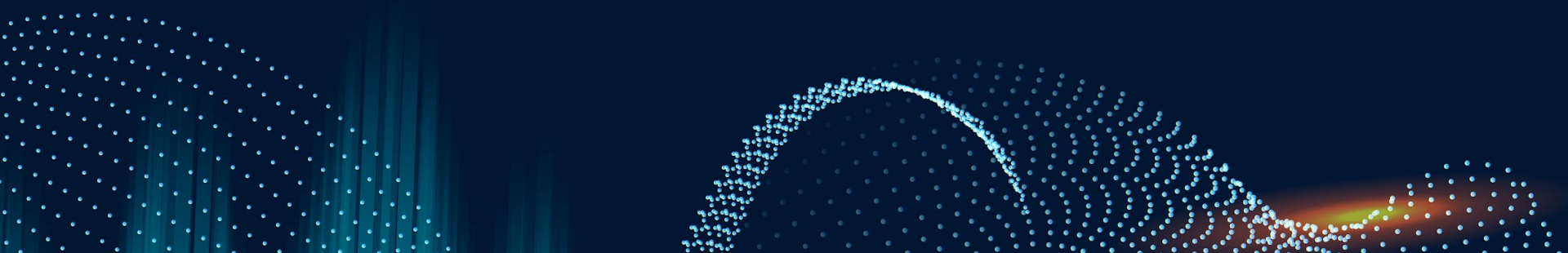
- Date of closing → days since the earliest date of the training set
- Categorical variables → one-hot encoding if few categories
- Categorical variables → group by cumulative frequency → one-hot encoding of the groups

Missing Values

- Latitude and Longitude → mean lat and long of the city
- Living Area → mean living area for the total number of rooms
- Other → median, zero, remove

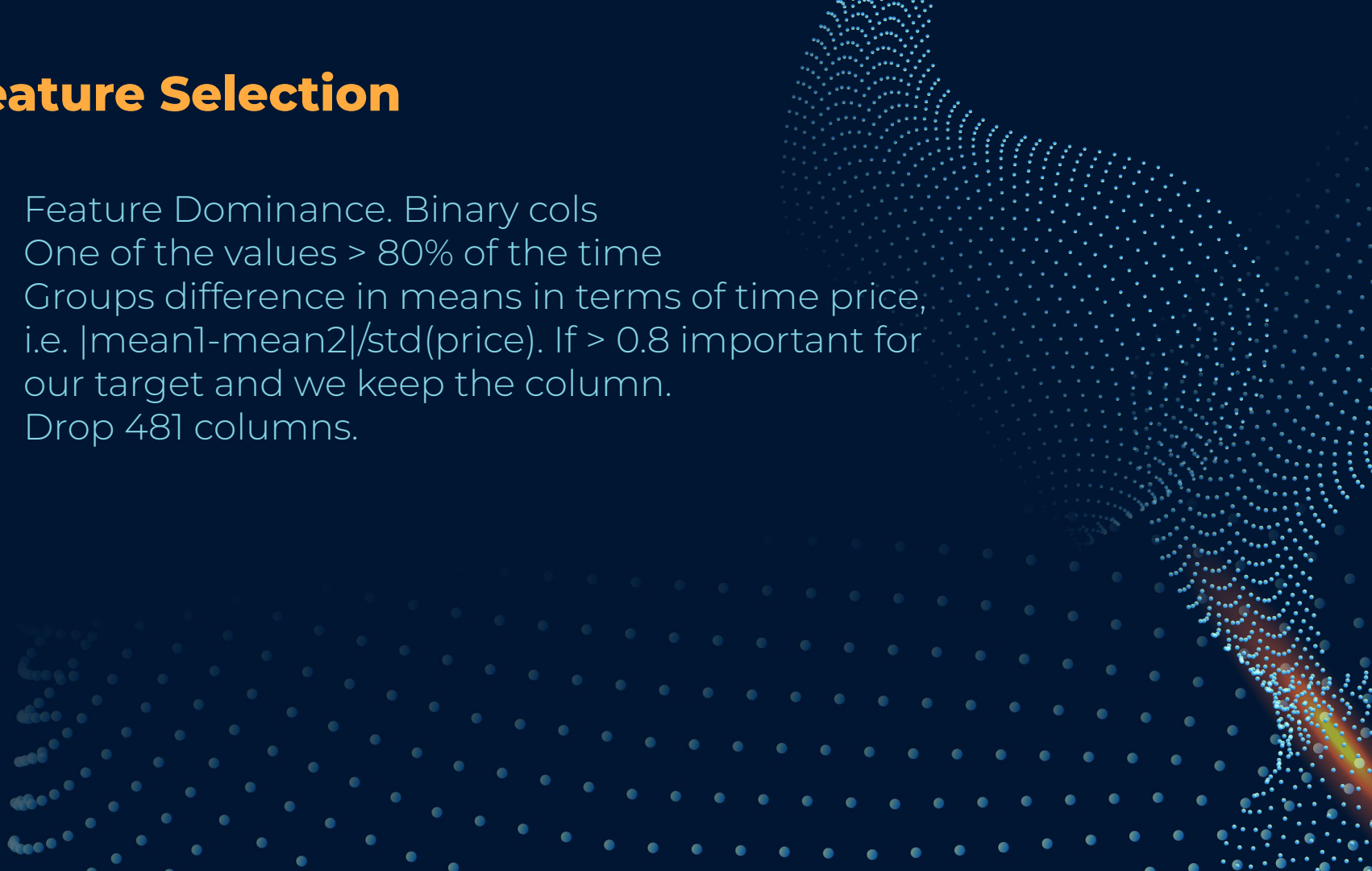
Outliers

- Search for peculiarities like:
  - 35M $ two-bedroom, single bathroom property
  - 75 bathrooms
  - 21 fireplaces
  - 20+ garage spaces
  - Properties built before 1800



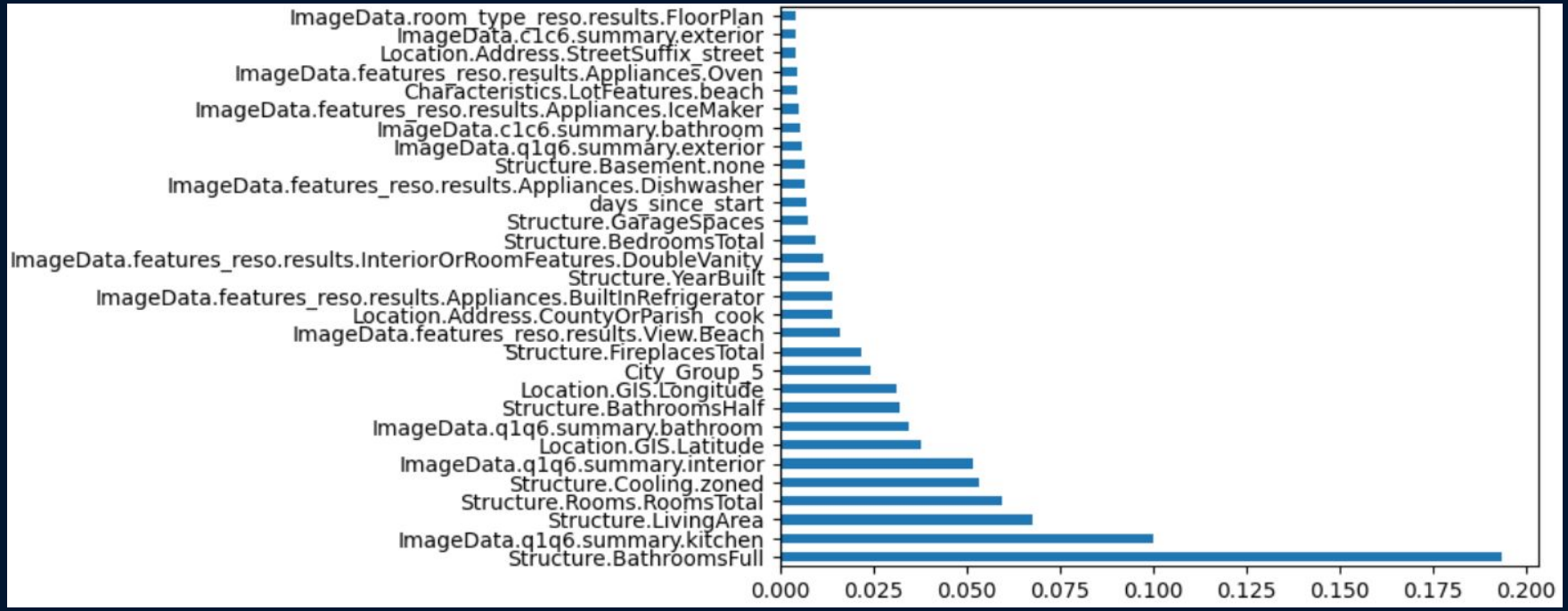
Feature Selection

- Feature Dominance. Binary cols
- One of the values > 80% of the time
- Groups difference in means in terms of time price, i.e. |mean1-mean2|/std(price). If > 0.8 important for our target and we keep the column.
- Drop 481 columns.



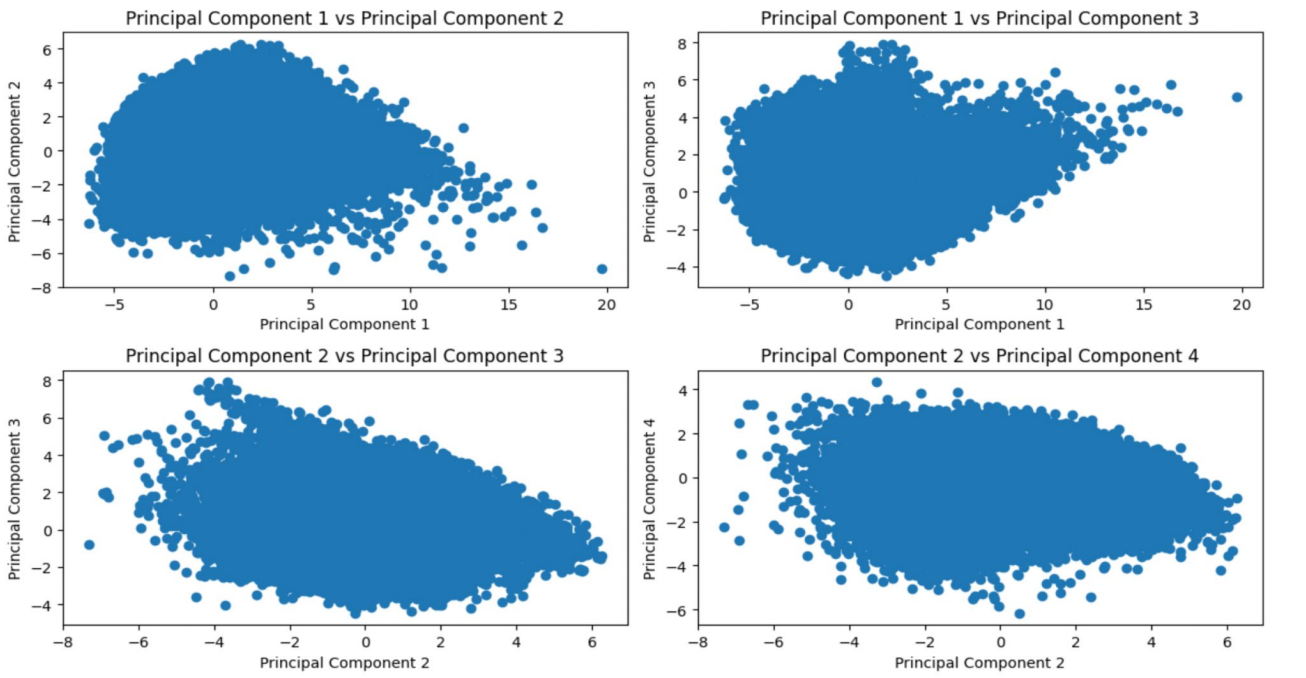
Feature Importance

- High correlated features (>0.8)
- Feature Importance - ExtraTrees. Drop unimportant features.



Feature Importance

- PCA. Similar results
- 24 feats kept



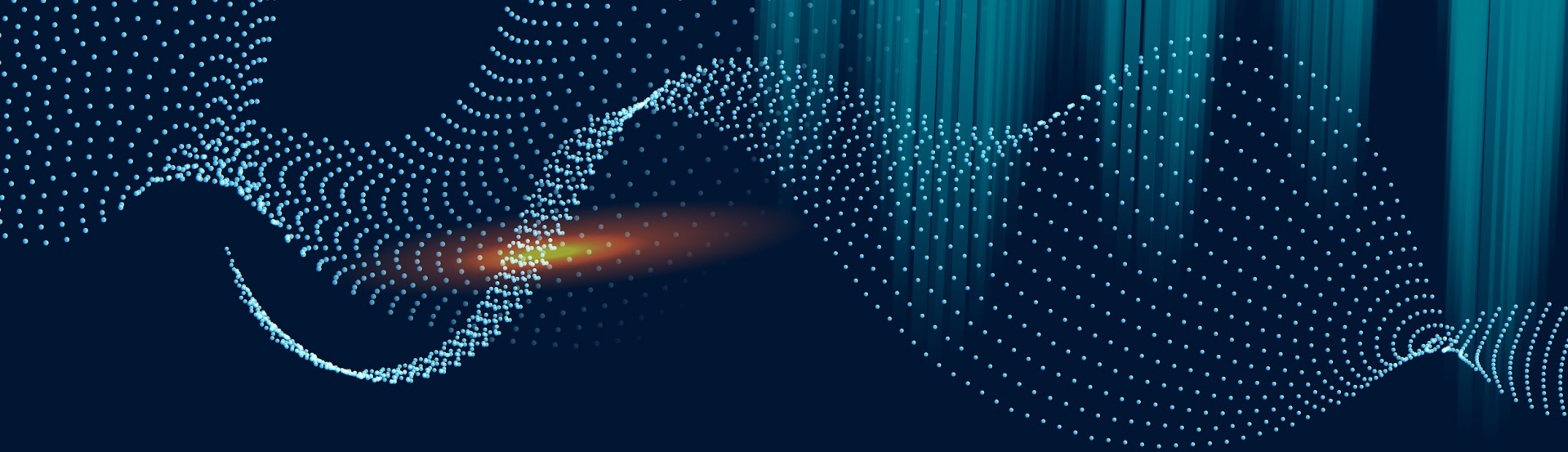
Feature Importance

- **Number of Bathrooms**
- **Kitchen quality from ImageData**
- Living area
- Number of rooms
- Zoned cooling system
- Interiors quality from ImageData
- Latitude

Modeling

- Split train val sets (0.2)
- Grid Search in train set
- ML algorithms tried:
  - LinearRegression
  - DecisionTrees
  - RandomForest
  - ExtraTrees
  - XGBoost

- Best Model: **RandomForest**(max_depth=30, estimators=200)
- MSE, MAE, MAPE
- Refit in train + val
- Validation set MAE: 58.000
- Preds Test set MAE: 80.000



Thank you for your attention!

Andreja Andrejic
Jinheng Lin
Simone Paloschi
Alexandros Tremopoulos