Apartment Price Prediction Model Report

1. Modeling Method

Target Variable

The model predicts **Price Per Square Meter** rather than the total price.

- **Justification**: Real estate prices are highly correlated with area. Predicting price/m² normalizes for size, reducing variance and allowing the model to focus on layout and quality characteristics (floor, location, finishing). Total price is derived as `PricePerMeter * TotalArea`.

Algorithm

Gradient Boosting Regressor (sklearn) was selected.

- **Justification**: Tree-based ensembles handle tabular data with categorical features (converted via target encoding) and non-linear relationships effectively without extensive scaling. They are robust to outliers and provide feature importance for interpretability.
- **Hyperparameters**: Optimized for generalization (`max_depth=7`, `n_estimators=2000`, `learning_rate=0.02`, `subsample=0.8`).

Feature Engineering

Key features driving the model:

- **Layout**: Rooms, Area (Total/Living/Kitchen), ratios (Kitchen-to-Total).
- **Floor**: Floor number, total floors, relative position (First/Middle/Top).
- **Location/Building**: District, Class (Economy/Business), Building Type (Monolith/Panel).
- **Derived**: Polynomial interactions (e.g., `TotalArea^2`) and Target Encodings (mean/median price per category) were crucial for capturing location value.

2. Validation Strategy

Train-Test Split

A **Stratified Shuffle Split** strategy was used to separate data into:

- **Training Set**: 85%
- **Test Set**: 15%
- **Stratification**: Done specifically on **Price Bins**. This ensures the training and test sets have statistically identical distributions of the target variable, preventing bias where the test set might contain only "expensive" or "cheap" apartments.

Data Leakage Prevention

Target encoding was performed **after** the split. Statistics (mean/median price by district) were calculated solely on the **Training Set** and mapped to the Test Set. Unseen categories in the test set were imputed using the global training mean.

3. Performance Metrics

The model was evaluated using:

Metric	Training	Testing	Interpretation
RMSE	~17.9k	~23.3k	Root Mean Squared Error. Penalizes large errors. The gap suggests some overfitting but acceptable for real-world variance.
MAE	~11.8k	~14.6k	Mean Absolute Error. On average, the prediction is off by ~14,600 ■/m^2.

R^2	0.93	0.82	Explains 82% of the price variance on unseen data.

Real-world Accuracy: Over **80% of predictions fall within a 5% error margin** of the actual price, making the model practically useful for estimation.

4. Assumptions and Limitations

Assumptions

- **Market Stability**: The model assumes current market conditions hold; typically, it does not account for temporal inflation unless "Listing Date" is explicitly modeled (currently static snapshot).
- **Data Quality**: Relies on accurate manual inputs. "Years to Handover" assumes linear depreciation/appreciation.

Limitations

- **Overfitting**: There is a ~23% drop in performance from Train to Test RMSE, indicating the model learns some noise.
- **Unseen Locations**: Districts or complexes not present in the training set will default to the global average, potentially leading to high errors for unique new developments.
- **Link Parsing**: The current link-based input is a prototype (mock). Production use requires a live scraper for `samolet.ru`.