HUSGRUPPEN: Cost Estimation Project Final Report

Group: mlproject — Machine Learning TMLS22 — Jönköping University

Eduardo Pereira, Juan Carlos Gonzalez Aguilar,
Julian Emery, Kezia Aprilia Sanjaya,
Muhammad Malik

***Abstract*—This study addresses the challenge of automating cost estimation for Husgruppen AB, a construction company that currently relies on manual processes that are both time-consuming and prone to human error. We developed a machine learning solution capable of predicting building costs using key construction parameters.**

Due to the limited size of our dataset (22 samples), we applied feature selection and data augmentation strategies to improve generalization. Synthetic samples were generated with controlled Gaussian noise to maintain statistical integrity. We evaluated several regression models, including Linear Regression, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP), using Mean Absolute Error (MAE) and R^2 as performance metrics.

Although the Multilayer Perceptron achieved slightly better results, we selected the Linear Regression model due to its strong balance between accuracy (MAE: 113,459 SEK, R^2: 0.8363) and interpretability. Our findings identified the snow zone, number of sliding doors, number of windows, wall area, and number of doors as the most significant cost predictors.

The final model offers Husgruppen AB a robust, easy-to-implement tool that improves operational efficiency and enhances customer experience while delivering valuable insights into key cost drivers.

I. Introduction

Husgruppen AB currently uses a manual approach to estimate building costs, a process that is both inefficient and vulnerable to human error. This traditional method results in inconsistencies and increased processing time, impacting customer satisfaction and internal workflows.

To address this, we propose an AI-powered cost estimation system that automates the process by accepting key input parameters, such as wall area, roof area, number of doors and sliding doors, geographic location, and snow zone, and generating real-time estimates. This solution promises faster, more accurate, and more consistent cost assessments.

The machine learning model incorporates the following construction parameters: wall area (the total surface area of all walls in the building), roof area (the total roof area, which affects material and labor costs), doors, windows, location and snow zone.

By integrating these parameters, the system provides a scalable and user-friendly interface for generating accurate cost estimates, supporting more informed decision-making for both Husgruppen and its customers.

II. Related Work

Habib et al. (2025) introduced an ensemble learning framework combining regression random forests and gradient-boosting regression trees that demonstrated significant performance improvements over support vector regression when applied to San Francisco's building inspection dataset. Their research highlights how ensemble methods can address the construction industry's challenge of limited data availability, as "the construction domain suffers from having datasets with limited data size, which prevents us from using deep learning techniques." While their approach focused on larger commercial projects, our work addresses the specific needs of smaller regional construction firms with extremely limited historical data.

Construction cost estimation challenges stem from numerous factors documented in recent literature. As Habib et al. note, "98 percent of mega projects suffer cost overruns of more than 30 percent," with traditional methods suffering from "time consumption, lack of previous cost information, market's fluctuating states, lack of resources, estimator's incompetency, inflation, and lack of construction knowledge."

Our research directly addresses these limitations by developing an automated system for Husgruppen AB that reduces human bias while providing consistent estimates based on key construction parameters relevant to the Swedish market. Unlike previous research focused on government datasets, our approach demonstrates that even private companies with very small datasets can implement effective machine learning solutions through appropriate data augmentation and feature selection techniques.

Our data augmentation strategy draws inspiration from Shmuel et al. (2025), who proposed using machine learning

models to generate synthetic data for improving deep learning performance on tabular regression tasks. Similar to their approach, we introduce Gaussian noise to key features in the original data and use a trained regression model to predict the corresponding target values for the augmented samples. This ensures that the statistical relationships within the data are preserved. As demonstrated in the referenced study, this method enhances model generalization and performance—especially when original data is limited—by expanding the dataset with realistic, model-informed samples. We adopt this principle to enrich our villa pricing dataset, confirming through distribution analysis and performance metrics that our augmentation maintains natural variability while boosting learning outcomes.

III. Background

This section outlines the core methods and concepts used throughout the project, focusing on those relevant to our approach.

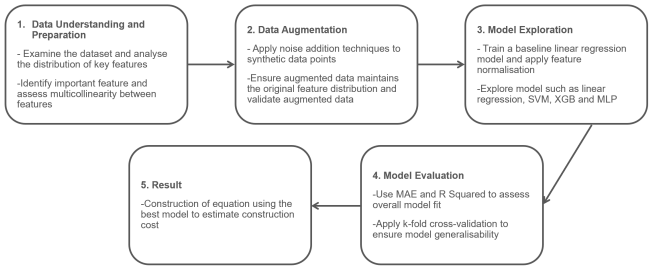

Fig. 1. Model Evaluation Workflow

- Regression Modeling: We explored multiple regression techniques, including Linear Regression, Support Vector Machines (SVM), XGBoost, and Multilayer Perceptron (MLP), to predict cost values from input features.

- Feature Selection: To improve model efficiency and interpretability, we employed a feature selection process that identifies the most relevant inputs for the prediction task.

- Data Normalization: Before training, we applied min-max normalization to scale features, enhancing model performance—particularly for algorithms sensitive to feature magnitude like Linear Regression and MLP.

- Data Augmentation: Given the limited size of our dataset, we used controlled noise to create synthetic samples. These were then used to train models on larger, more diverse datasets without overfitting.

- Evaluation Metrics: We evaluated model performance using Mean Absolute Error (MAE) for its practical interpretability, and R^2 as a complementary measure of explanatory power.

IV. Data Exploration

The dataset provided for this project contains key architectural and pricing attributes related to house structures, with the objective of predicting the final house price (including tax). The data includes physical measurements such as surface areas, the number and types of openings (windows and doors), the snow zone of the location, and delivery-related information. However, we were instructed to ignore the delivery location, as the focus of the project is to estimate the price of the house itself, excluding any delivery-related costs.

A. Overview of the Features

Feature	Description
Wall area (m²)	External wall surface area
Inner walls area (m²)	Internal wall surface area
Roof area (m²)	Total area covered by the roof
Floor area 2nd floor (m²)	Area of the second floor
Delivery location	Categorical feature indicating destination region (ignored)
Number of windows (std size)	Count of standard-sized windows
Number of windows (¿18 height)	Count of large windows (over 18 in height)
Number of sliding doors	Count of sliding doors
Number of window doors	Count of doors with window panels
Number of doors	Count of standard doors
Price 6-10 (SEK excl. tax)	Price of all windows and doors, excluding tax
Snow zone	Numerical indicator of the regional snow load
Price incl. tax	Final quoted price of the house, including tax

TABLE I
FEATURE DESCRIPTIONS IN THE DATASET

B. Initial Observations

From the initial analysis, it became clear that certain features would have a stronger influence on the house price than others. For example, the wall area and the number of openings (windows and doors) are likely to contribute significantly to the final cost. In contrast, features like inner wall area or delivery location (which was excluded) appeared to have less direct impact.

The price values in the dataset followed a relatively intuitive pattern: the greater the number of windows and doors, the higher the overall cost. This provided an early indication of the direction we would take in our modeling and feature engineering.

C. Dataset Quality and Size

The dataset was well-structured, with no missing values or formatting issues across any features. This made the initial preprocessing phase efficient, allowing us to proceed directly to analysis and feature engineering.

However, a major limitation was the small dataset size, consisting of only 22 samples. This limited sample size introduced challenges in building robust models and increased the risk of overfitting. With so few data points, the model's ability to generalize to unseen data is reduced, and caution must be taken when interpreting patterns or evaluating feature importance.

D. Correlation Analysis

To better understand the relationships between numerical features in the dataset, we computed and visualized a correlation matrix. This analysis provided valuable insights into which features are most strongly associated with the target variable, *Price (SEK exkl tax) incl 6-10*.

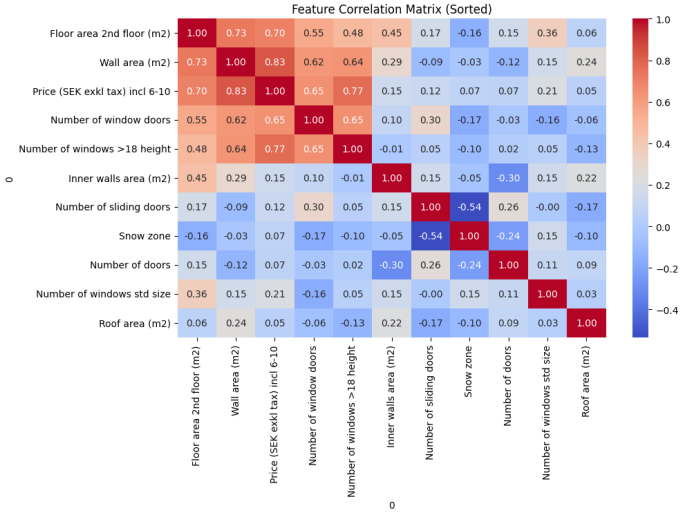

Fig. 2. Feature Correlation Matrix (Sorted)

From the heatmap, we observe that **Wall area (m²)** has the highest positive correlation with the target price (correlation coefficient of 0.83), followed by **Number of windows $>$ 18 height** (0.77) and **Floor area 2nd floor (m²)** (0.70). These features appear to be key drivers of house price, likely due to their influence on material and construction costs.

Interestingly, some features like **Snow zone** and **Number of standard-sized windows** showed weak or even slightly negative correlations with price, contrary to initial expectations. This suggests that either their pricing impact is less direct or already embedded in other more influential features.

Additionally, features such as **Number of sliding doors** and **Number of doors** had low correlation values with the target variable, indicating they may have minimal predictive power on their own. This is mostly explained by the fact that most of the houses in the datasets do not or have a very few amount of them, thus making it hard to find any correlation, but these features remain important to estimate correctly the price, for potential future samples.

Overall, this analysis guided our feature selection and engineering strategies, allowing us to prioritize features with stronger relationships to the price.

V. Approach Description

Based on the dataset size (22 rows) and the amount of the features, the approach was implemented using a feature-selection step to find the best set of features for our final models. Then it was used a crossvalidation step of 10 folds, it was decided to go with 10 because in this way we ensure the amount of rows used for training is between 19-20 and the validation rows are between 2-3. Since for each step, we are validating in unseen rows this is a way to validate that our approach is generalizing the problem and we are not overfitting.

As a preparation of all the features, it was performed a min-max normalization, to ensure the Linear Regression and Multilayer Perceptron performed well.

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{1}$$

A. Data Augmentation

Our data augmentation strategy focuses on enriching a villa pricing dataset by generating synthetic samples that maintain realistic statistical properties. The core idea involves selecting important numerical features (e.g., area, number of rooms, distance to amenities) and introducing low-level, feature-specific noise to these values. For instance, only 2–5 percent Gaussian noise is added depending on the feature type, ensuring that the augmented samples remain plausible. We use a trained Random Forest Regressor on the original dataset to estimate the corresponding prices of these noisy samples, and slight perturbations (1 percent noise) are also applied to the predicted prices. These augmented entries are then appended to the dataset.

To ensure the usefulness and validity of the augmented data, we conduct distribution comparisons and performance evaluations. We compare histograms and kernel density estimates (KDE) for each feature between the original and augmented datasets to verify that the statistical distributions remain consistent. Furthermore, we evaluate the model's predictive performance with different amounts of augmented data added and assess the performance using both RMSE and R² metrics.

B. Model Evaluation

For the model evaluation, linear regression, Multilayer Perceptron, Support Vector Machine (SVM), and XGBoost were used to evaluate various model families and find the one that best fits our problem. Using the results of the feature selection step we trained all the models with data-augmented training sets and evaluate them using real samples that were not used during the data-augmentation step to ensure not giving to the model extra information that call result in overfitting or false results.

Since for the business is more important to have predictions similar to the real ones we chose Mean Absolute Error (MAE) as the main metric to measure the performance of the models. However, we used R^2 as a complementary metric.

The feature selection process was carried out by testing all possible combinations of features. This approach was chosen due to the small number of available features and samples. We evaluated combinations ranging from 1 to 10 features. For this process, we used the `Sequential Feature Selector`,

which iteratively adds the feature that most improves the current subset. The selection continues until the desired number of features is reached.

To have a wider test of the features we used cross-validation and the model mentioned for the model evaluation, this process is started early to the full model evaluation.

VI. Experiment Design

For the experiment design, as it was mentioned in the section. V-B we used 10-folds in the cross-validation step. After that, we perform data augmentation in 9-folds and test without seeing 1 remaining fold. With this design, we guarantee more reliable performance metrics. The size of the data augmentation was provided using the results of the data-augmentation design (see section VII-A).

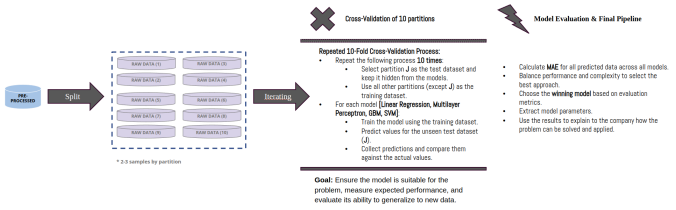

Fig. 3. Model Evaluation Workflow

During the model selection part it was defined the hyp-parameters used on each model (see table II) based on the nature of the problem. In particular "lbfgs" because in fast-testing we found was the best-performed metrics. For SVM was choose degree of 2 because linearity is tested using Linear Regression. For XGB we used only 2 estimators because of the size of the training set. Basic setups were defined to do hyper-parameter tuning with the winner model.

As a baseline, it was used the average of all the training set prices to predict the price of the test dataset.

Model	Hyperparameter	Value
Baseline	–	–
LinearRegression (without oversampling)	–	–
LinearRegression	–	–
SVM (SVR)	kernel degree C	poly 2 1.0
XGBRegressor	objective n_estimators random_state device	reg:absoluteerror 2 SEED cpu
MLPRegressor	hidden_layer_sizes activation solver alpha max_iter learning_rate random_state validation_fraction	(100) relu lbfgs 0 10000 adaptive SEED 0.3

TABLE II
Models and Their Selected Hyperparameters

It should be mentioned that it was defined as a seed the value 12345 to ensure reproducibility.

VII. Result Analysis

A. Data Augmentation Results

We evaluate the model's performance using datasets augmented with varying sample sizes (100, 200, 500, 1000).

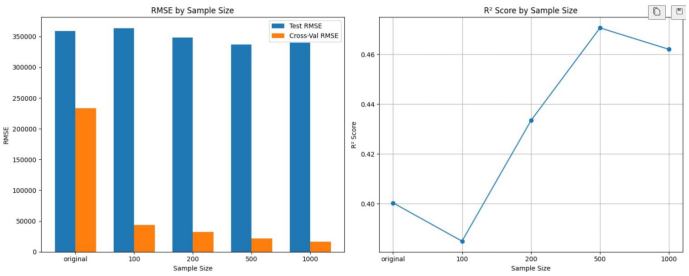

Fig. 4. Data Augmentation Evaluation Using Varying Sample Size

From fig.3., we observe that with an augmentation of 100 samples, the RMSE increases and R^2 drops. This is likely because the model is still learning to adapt to the new synthetic data and has not yet generalized effectively. As the augmentation increases to 200 and 500 samples, the RMSE steadily decreases and R^2 improves, indicating that the model benefits from the added variability and begins to learn more robust patterns. However, when the sample size exceeds 500, performance slightly declines—suggesting that the model may start memorizing noise in the augmented data rather than generalizing from it. This trend highlights 500 as the optimal number of augmented samples, offering the best trade-off between data enrichment and model generalization.

Fig.4. shows that the augmented features (orange) closely resemble the original data (blue), with overlapping patterns and similar density curves across key features like wall area, inner walls, and roof area. This similarity is crucial to maintain natural variability, ensuring that the synthetic data enhances model learning without introducing unrealistic patterns or bias.

B. Feature Selection Results

After running the feature selection search we found that the best combination of features is using 5 (see Figure 6).

The features selected are: `['wall_area_m2', 'number_of_windows_gre18_height', 'number_sliding_doors', 'number_doors', 'snow_zone']`.

C. Model Selection Results

The result of the min-max normalization can bee seen in the table III.

After the 10 executions in the cross-validation process the results we decided to go for the Linear Regression model. Although the best model using the metrics was the Multilayer Perceptron (MLP) it was decided to go with the Linear

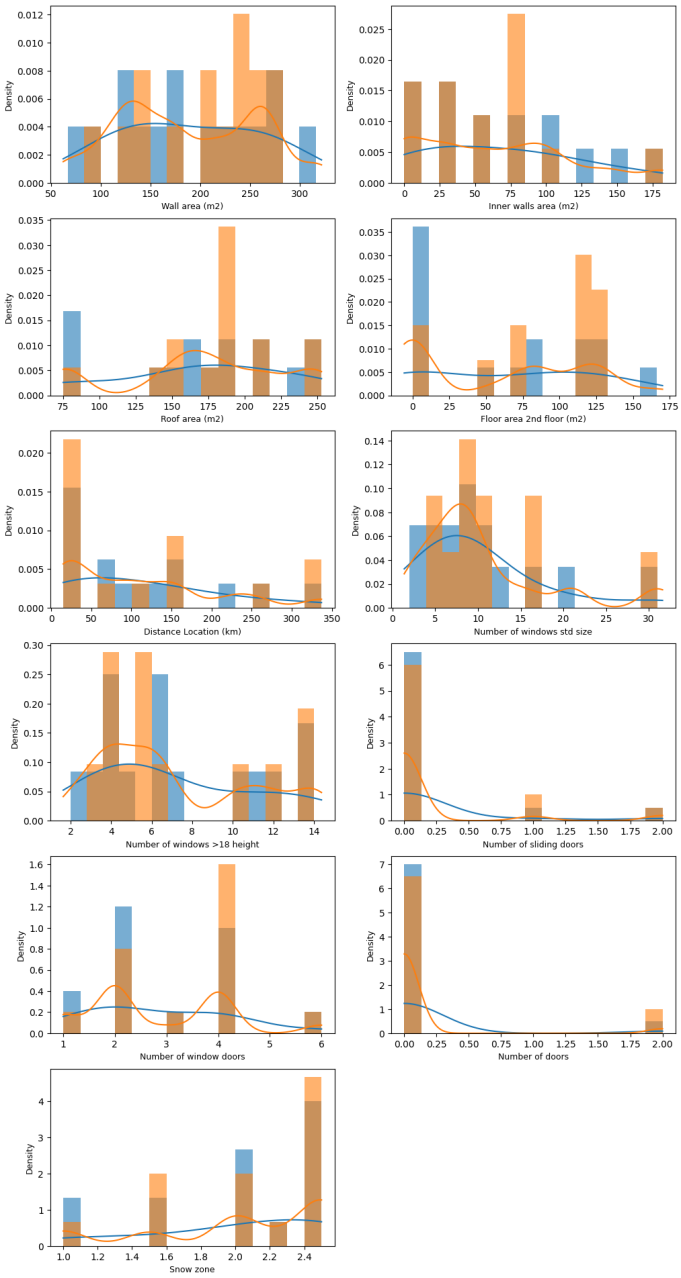

Fig. 5. Comparison of Data Distribution Between Original and Augmented Feature

Feature	Min	Max
wall_area_m2	67.0	316.0
number_of_windows_gre18_height	1.0	22.0
number_sliding_doors	0.0	2.0
number_doors	0.0	2.0
snow_zone	1.0	2.5

TABLE III
MINIMUM AND MAXIMUM VALUES OF SELECTED FEATURES USED IN THE MODEL.

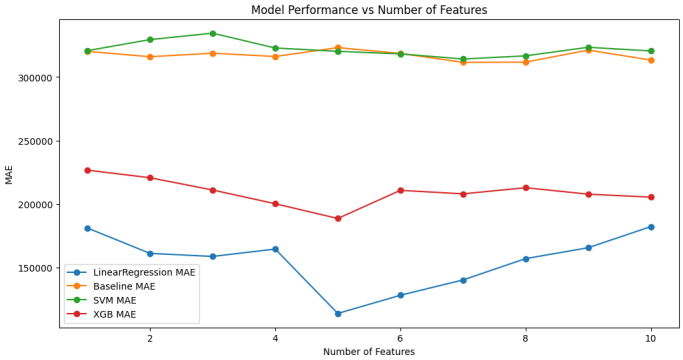

Fig. 6. Feature selection: This feature selection process was ran using Sequential Selector.

	LR*	BS	XGB	LR	MLP	SVM
MAE	114,234	312,418	190,478	113,459	112,948	310,342
R^2	0.8360	-0.0725	0.4818	0.8363	0.8372	-0.2137

TABLE IV
PERFORMANCE METRICS (MAE IN KR AND R^2) FOR EACH MODEL. LR STANDS FOR LINEAR REGRESSION, LR* FOR LINEAR REGRESSION NOT USING DATA AUGMENTATION. THE VALUES FOR MAE WERE ROUNDED TO THE CLOSEST INTEGER SINCE ALL THE TARGETS WERE INTEGERS AND FOR VISUALIZATION PURPOSES. ADDITIONALLY, R^2 WAS ROUNDED TO 4 DIGITS. NOTICE THAT MULTILAYER PERCEPTRON WAS THE MODEL THAT PERFORMED BETTER. IN SECOND PLACE IS THE LINEAR REGRESSION USING DATA AUGMENTATION

Regression because is easier to explain to the final user and also is easier to implement in production.

Looking in figure 8, it is clear that our model can be used to have a good approximation of the real home building cost. In the other hand, there is evidence that the 20th observation is an outlier, in the future this observation should be ignored, or more features to differentiate it.

D. Final Pipeline

For the pipeline, the model was trained with the totality of the dataset. Then we got the weights for all the features chosen.

The weights of the normalized features are:

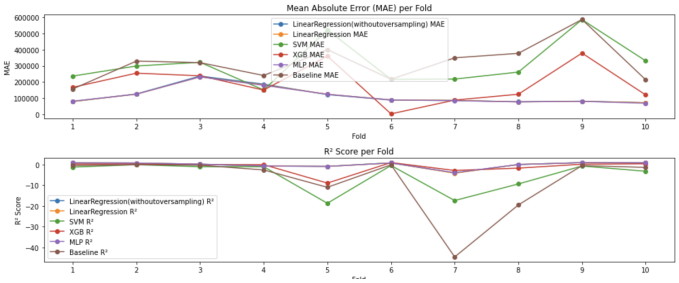

Fig. 7. Model Evaluation Performance Plot: The models based on a linear combination of the features were indistinguishable. However, they were the best performed overall, one hypothesis from the authors is with more samples XGB could perform almost as well or better than Linear Regression.

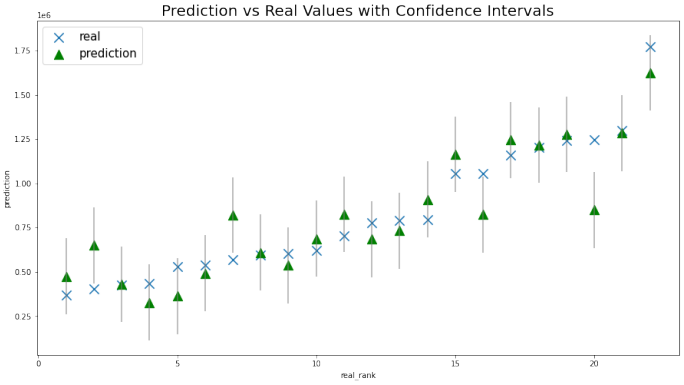

Fig. 8. Prediction vs Real values with confidence interval: In this plot, we can observe that the prediction predicts high values for real high cost and for lower as well, which means that the model can be used to prioritize quotations. Another thing to observe is that the confidence interval of one standard deviation (in this case 209,851 kr for each side) can be used to give a confidence interval having real values in the 81.82% of the times.

$$\begin{aligned} \text{price\_sek\_excl\_tax\_incl\_6\_10} = \\ 816151.28 \cdot \text{normalized\_wall\_area\_m2} + \\ 544436.32 \cdot \text{normalized\_number\_of\_windows\_gre18\_height} + \\ 295032.06 \cdot \text{normalized\_number\_sliding\_doors} + \\ 172596.66 \cdot \text{normalized\_number\_doors} + \\ 371516.66 \cdot \text{normalized\_snow\_zone} - 24079.20 \end{aligned} \tag{2}$$

To de-normalize we will separate individual features and get the real weights getting an equivalent equation:

$$\begin{aligned} \text{price\_sek\_excl\_tax\_incl\_6\_10} = \\ -517289.49 + \\ 3277.72 \cdot \text{wall\_area\_m2} + \\ 25925.54 \cdot \text{number\_of\_windows\_gre18\_height} + \\ 147516.03 \cdot \text{number\_sliding\_doors} + \\ 86298.33 \cdot \text{number\_doors} + \\ 247677.77 \cdot \text{snow\_zone} \end{aligned} \tag{3}$$

Using this equation we can expect for example that for each additional square meter in the wall area we can expect to have 3,378 krs more expensive cost in the final quotation (see Table V). This information is useful for the company to have a better understanding of how their features affect the final cost.

Feature	**Weight (Coefficient)**
wall_area_m2	3,277.72
number_of_windows_gre18_height	25,925.54
number_sliding_doors	147,516.03
number_doors	86,298.33
snow_zone	247,677.77
Bias (Intercept)	-517,289.49

TABLE V
RAW COEFFICIENTS OF THE LINEAR REGRESSION MODEL PREDICTING PRICE_SEK_EXCL_TAX_INCL_6_10.

VIII. CONCLUSION

As a result of all the work done in this work the first conclusion found that the building cost is linearly correlated to the fields: `['wall_area_m2', 'number_of_windows_gre18_height', 'number_sliding_doors', 'number_doors', 'snow_zone']`. Additionally, the feature that most influences the total cost is the snow zone, followed by the number of sliding doors. The first can be explained because difficult conditions need more expensive materials in the house, and the sliding house can be correlated with a more complex and expensive building.

Secondly, we demonstrate that overall linear-based models have the best performances and that having a large amount of samples will improve the quality of the model, demonstrated with the improvement of the model using data augmentation.

Thirdly, it was demonstrated that it is possible to have a simple statistical model using Linear Regression that can be used for the company to deliver an easy-to-implement tool that can help save cost on quotations focusing on more expensive projects and knowing the feature that most impact the final cost, a recommendation can be to get more features related to the most important features, for example, more values or more detailed values for snow zone.

Finally, this project demonstrates that it is possible to train machine learning models using the company's own data. This implies that the company could benefit from collecting more features and expanding its datasets, which could enhance internal processes and support data-driven sales strategies.

IX. REFERENCES

Habib, O., Abouhamad, M., Bayoumi, A. (2024). Ensemble learning framework for forecasting construction costs. Automation in Construction, 170, 105903. https://doi.org/10.1016/j.autcon.2024.105903

Shmuel, A., Glickman, O., Lazenbik, T. (2025). Data Augmentation for Deep Learning Regression Tasks by Machine Learning Models. https://arxiv.org/html/2501.03654v1