The effect of immigrant demographic on housing prices, a case study of Dalarna county, Sweden

Employing Supervised Machine Learning Models for Analysis and Prediction

Joycelyn Laryea, Julateh K.Mulbah, Rupasinghe J. M. Nipunika C. Jayasundara
Business Intelligence
Dalarna University
h19joyla@du.se
v19julmu@du.se
h18rupja@du.se

***Abstract*—In this study, the empirical evidence of building a data driven model to predict prices of houses in Dalarna County was examined. The focus was on the effect of the immigrant population in determining the price of houses. This project used housing price data from hemnet.se for the period 2013-2019, immigrant data for the related period was sourced from Statistics Sweden. The employed supervised machine learning methodologies learn the patterns of the features within the training dataset and based on that predict housing prices for a dataset yet unseen. The random forest algorithm gave the best model with an accuracy of 56%. However, it was observed that of all the features, the migrant population was the least influential in determining the prices of houses in Dalarna county.**

Keywords-component; machine learning, immigrants, Dalarna, linear regression, decision tree, random forest.

I. INTRODUCTION

Depending on the country status, there are several reasons to invest in a house. People who live in developing and under developing countries invest in a house or simply purchase a house to satisfy their basic necessity of shelter. On the other hand, purchasing a house is one of the major investment plans of many people who live in a developed country like Sweden.

There are many determinants of the housing price changes in Sweden. Those determinants are housing attributes such as number of rooms and number of square meters, location of the house and some economic factors such as the income of people, housing loan interest rate, inflation rate and supply and demand for houses [1], [2], [3].

The cyclical nature of the Swedish housing market has been documented in academic literature. There is much research done regarding housing price changes in Sweden's urban cities such as Stockholm, Malmö and Gothenburg based on economic factors [4], [5].

The observed research gap in the academic literature is the lack of studies in housing price changes in underpopulated or rural areas in Sweden, as well as the effect of some social factors such as the number of immigrants. To help address this research gap, this study investigates the effect of the immigrant population on housing prices in Dalarna county and predict prices given foreseeable immigrant numbers to aid investors in decision making.

The result of this project can support the Dalarna region and the different municipalities in Dalarna county when locating immigrants and providing places of residence for them. As well, housing companies in Dalarna such as Tunabyggen and HSB homes can use the information and identify housing price trends.

II. LITERATURE REVIEW

There is vast literature that explores the determinants of housing prices and these literatures tend to focus on the North American, Europe and South East Asian markets. According to the available literature, the determinants of housing prices vary; from the interior determinants to the external determinants.

Internal determinants are interior design and the location of the house. External determinants are demographic factors such as age, income level of investors, population growth etc. and macro and micro economic factors such as housing loan interest rates, housing policies etc. [6].

Many researchers have left unaddressed a crucial demographic factor which is the increasing level of immigrants when determining housing price changes. According to Statistics Sweden, total immigration to Sweden for 2017 was expected to be roughly 180,000 individuals, and thereafter to 110,000 persons every year.

The empirical findings by Kenneth state that the internal determinants can be categorized as property determinants, community determinants and proximity determinants. Property determinants are types of houses and the number of rooms available in the house and by increasing the room numbers, housing price is increased [7].

Sweden's housing market has several types of houses. In regards to family living, there are 3 main types of houses available; Villa, Lägenhet (Apartments) and Radhus (Townhouse or Chain house). Villa type houses are more

expensive than Lägenhet and Radhus. However, sometimes it can differ according to the location [1].

Community determinants and proximity determinants are the population in a selected area and access to that particular area such as availability of motorways and distance from the city center. According to Kenneth's empirical findings, regions with high population and community facilities have high house prices.

From the above, we generate the hypothesis for this work i.e. Immigrant numbers in municipalities (Kommuns), show a positive impact on the housing price. Immigrant here in refers to people of other nationality besides Swedish.

III. MACHINE LEARNING MODELS

Machine learning refers to algorithms that learn from data in a recursive manner to identify trends, patterns and correlations from the dataset. Supervised machine learning is mostly applicable to datasets where past observations are a good predictor of the future. In this project, we used supervised machine learning algorithms with emphasis on regression since our dependent variable is quantitative [8].

Selected Methods - The following models were experimented with to provide insights for our hypothesis and led to our findings.

Multiple Linear Regression - This is an upgraded form of a simple linear regression because it accommodates more than one predictor variable. This means the model is expected to fit a relationship between a numerical outcome variable and a set of predictors. Below is the multiple linear Regression equation:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon\epsilon$$

Y represents the dependent variable here and e is the error or the unexplained part of the data.

$$\beta_o \ldots\ldots, \beta_p \; are \; cofficients$$

The multiple Linear Regression model does well with linearly separability phenomena and has faster computational capability but nevertheless it does not perform well in non-linearly-separable cases [8].

Decision Tree- This supervised machine learning technique helps to split the data nodes based on one particular variable. The aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The major advantage of the Decision Tree model is that it makes no assumptions about the relation between outcome and predictors and is also easier to explain as compared to other models. The disadvantage of this model is that it tends to overfit data, before a stop criterion is reached, it keeps splitting nodes over and over, learning all the details of the training data [8].

Random Forest – This is an ensemble of decision trees, most of the time trained with the bagging method. One advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Random forests bring extra randomness into the model, when it is growing the trees. It also has the features importance component that measures the relative importance of each feature [8].

IV. METHODOLOGY

A. *Source of Data*

The Housing Price Dataset for Darlana was accessed from hemnet.se [9] for the period 2013-2019, Immigrant data for the related period was accessed from Statistics Sweden [10]. Statistics Sweden is a data-oriented site that has many different datasets as it relates to Sweden while hemnet.se is an ecommerce site that sells houses around Sweden.

B. *Structure of Dataset*

The Dataset used for this project contains 15976 observations with six features. Out of these, the Price feature was the target variable while the remaining five were the independent variables. Migrant population, which is the focus of this research was one of the independent variables.

Table1. Description of each variable in the Housing prices dataset

Variables	Description
Price	Full prices of homes purchased in Darlana County
Size	The number of rooms associated with a house
Type	The type of houses bought that could be a Villa, Radhus or Lagenhet
Migrant pop	The migrant's population in a community
Kommum	The communities considered in this dataset
Year	The year these data was generated

C. Data Pre-Processing

The Housing price data and that of the migrant population were cleaned and integrated with the aid of Microsoft Excel based on the "Kommun" and "Year". This resulted in a 6 feature .csv file which we imported into the Python work environment with the aid of the Pandas library. The features were of type integer, float as well as objects.

	Year	Kommun	Price	Size	Type	Migrant_pop
0	2019	Falun	1360000	3.0	Lägenhet	3109
1	2019	Falun	1795000	5.0	Lägenhet	3109
2	2019	Falun	2410000	3.0	Lägenhet	3109
3	2019	Falun	880000	1.0	Lägenhet	3109
4	2019	Falun	1925000	3.0	Lägenhet	3109

Figure 1: Samples in dataset

With Pandas, Matplotlib and Seaborn libraries, we explored the dataset to understand the distribution of the features. Some observations were that, in the 7-year period, the year 2019 had the most sales activity. The housing type "Villa" saw the most transactions in our dataset. Three room houses were the most dominant and Borlange Kommun and Orsa were the most and least featured in our dataset respectively. This implies high and low levels of trading in houses in those Kommuns.

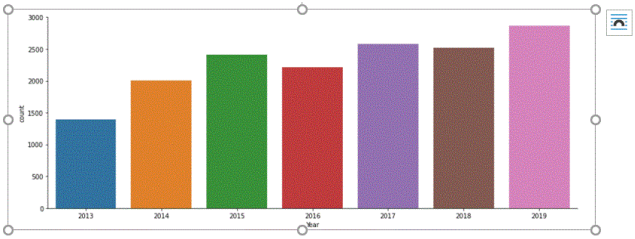

Figure 2: Year variable distribution

The year. 2016 recorded the highest migrant number by a Kommun at 4864 and this was in Borlange. 2017 in total had the most migrant numbers for the Dalarna county.

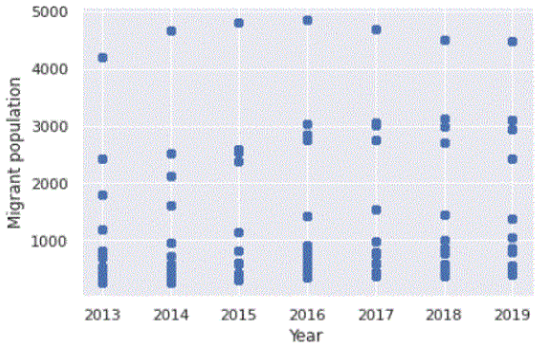

Figure 3: Immigrants population per year

Two columns (Size and Type) were observed with a total of 207 null values and these rows were dropped as we found them to be negligible in comparison to our over 15000 observations. The features Year, Kommun, Type and Size were then converted into categorical values, Migrant_pop to integer and Price to float. In the "Size" feature, it was observed that rooms of nine and above were outliers and thus a new data frame was created limiting our data by size to a maximum of eight.

During data exploration, it was also observed that quite a number of prices were outliers with a frequency of one, this was also confirmed with box plots showing some values above the upper quartile.

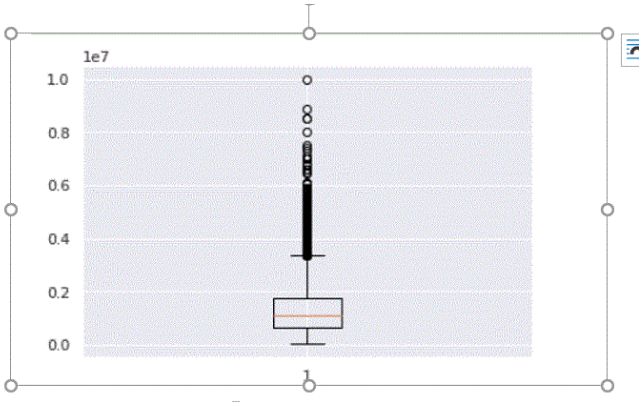

Figure 4: Price variable outliers and skewness towards the mean

We thus calculated the Z score which is the number of standard deviations by which a data point is above the mean. We set a threshold of 3 standard deviations above the mean to define what an outlier should be and went ahead to remove all outliers [11], [12]. This resulted in a dataset with 15422 observations of the 6 features.

A correlation analysis was administered to know if the independent variables share the same linear relationship with the housing price so as to detect duplication of variables in the dataset [13]. Fortunately, there was not much correlation among our independent variables. The following cluster of correlated features were found using Pearson correlation coefficient-r cutoff of (-0.6,0.6).

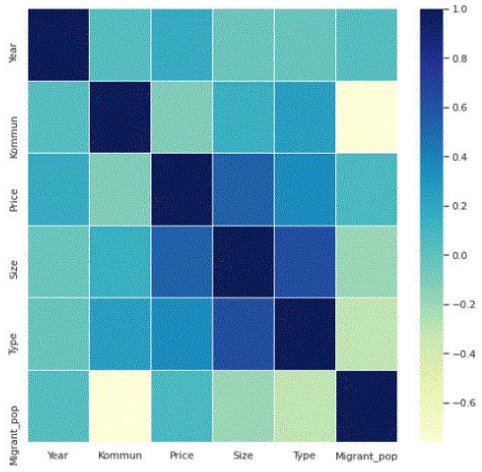

Figure 5: Heat map showing the correlation among the independent variable and dependent variable

For the purpose of our research, we proceeded to create a data frame with only the independent variables and took out Price

which is the dependent variable. To assess the relevance of each independent feature, we visualized the correlation between each feature and the dependent variable.

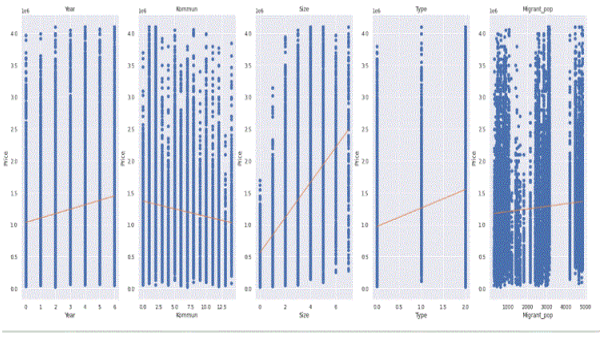

Figure 6: Relationship of price (dependent variable) with independent variables (full dataset)

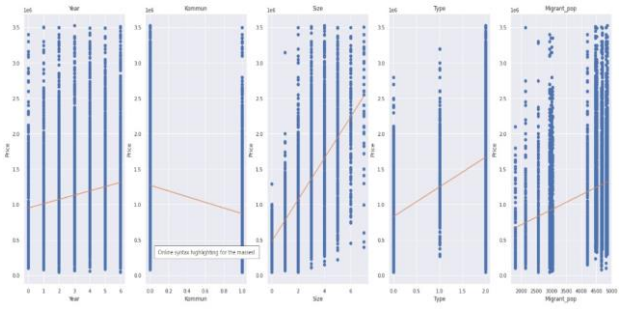

Figure 7: Relationship of price with independent variables (reduced dataset)

As would be expected a positive regression line was seen between the size feature and the price. However, though positive, out of all the independent variables, the feature of our hypothesis showed the least effect on the target variable. We investigated this further by filtering for the Kommuns with the most, minimal and average migrant numbers being Borlange, Vansbro and Ludvika. The migrant number seemed to show more of an influence in these 3 Kommuns but was even more relevant in the Kommun with the least migrant numbers.

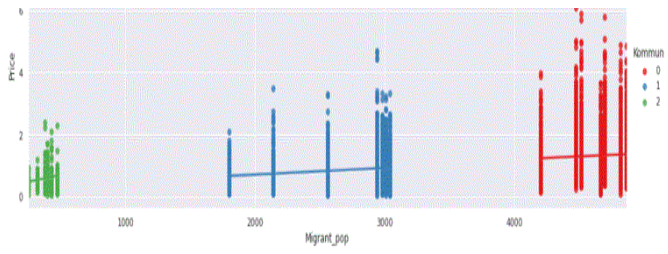

Figure 8: Relationship of Price with the migrant population for Vansbro, Ludvika and Borlänge kommuns

D. Modeling

Regression Decision Tree, Multiple Linear Regression and Random Forest methods were adopted to develop a predictive model from our dataset. These were done with the Sci-Kit Learn library.

Our data was partitioned at 80:20 for training and testing respectively. This was for our entire data set.

V. RESULTS

The coefficient of determination was the performance measure used to compare our regression models. Its values range from a worst of zero to a best of one and indicate how good the model is at making predictions. Denoted as R^2, it is the square of the correlation between the values predicted and the actual values and speaks to the proportion of the variance in the dependent variable that can be predicted by the independent variable(s).

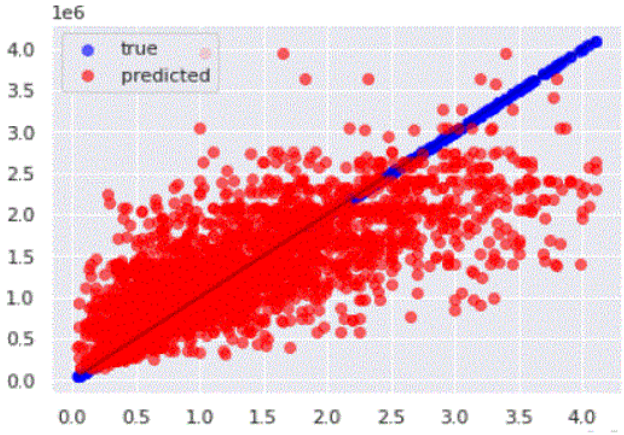

Figure 9: Decision tree result: Actual against predicted

In building a model with the decision tree, different maximum depths were experimented with, this indicates the number of levels that should be within the tree. The best model was at a depth of 10 for the full data set but at 5 for our reduced data set.

Our best model, though marginal, was with the random forest, which gave an accuracy of 0.56. The linear regression model was however the least accurate for our full dataset but produced the best prediction on the reduced dataset. Accuracy levels were relatively higher on our reduced data set, “Kommun” and the migrant population also showed stronger regression coefficient values in the reduced data set.

Table 2: Accuracy score for models on full-dataset

Model	
Random Forest	0.561590
Decision Tree	0.551053
Multiple Linear Regression	0.354643

Table 3: Accuracy score for models on reduced dataset

Model	Score
Decision Tree	0.626662
Random Forest	0.625374
Multiple Linear Regression	0.432538

VI. DISCUSSION

Through visualization and supervised learning algorithms, it has been realized that the migrant population is the least determinant of housing prices. However, with a reduced dataset, though it still remained the least relevant, it showed stronger correlation to the target variable, price. This could be an indication that the migrant population is more relevant in certain Kommuns than others and this is masked in the wider dataset. A Kommun level analysis could provide further insight. This is further corroborated by the linear regression model, though marginally, being the best fit for the reduced dataset whereas it was the worst for the overall dataset.

As expected, Size, Type and Year in that order were the most correlated to the target variable. It is noted that Kommun was the second least correlated independent variable to the dependent variable. In view of literature that posits a community or a location is significant in how much a house is sold for, the negative correlation observed between the Kommuns and housing price seem a contradiction. However, this could be as a result of the "positioning" of the Kommuns in the dataset, i.e. kommuns which are numbered at first ranks due to their alphabetical order have houses with higher selling prices compared to the kommuns which are numbered as last.

As mentioned in the literature review, the kommuns (e.g Falun, Borlänge) which received first numbers are high populated kommuns with sophisticated proximities such as availability of motorways, highways and city centers and have higher housing prices compared to the other kommuns. This explains the reason for the negative correlation as visualized. It does not contradict the literature that housing selling price and the location have a positive correlation.

The effect of outliers was also visible as initial work which had only taken out outliers with respect to the "Size" feature resulted in less accurate models. Not only was the predictive accuracy improved upon with the use of the z score to reduce the effect of outliers, but the correlation between the independent variables and the target variable was also enhanced.

This, together with the varying results observed with subsets of the data points at the likelihood of seeing a stronger correlation between the migrant population and the housing price should "noise" which in this instance refers to the effect of other Kommuns being removed.

VII. FUTURE WORK

Reference to the discussion above, a logical step for future work would be to investigate the effect of migrant population on individual Kommuns as it seems to be the case that literally one model does not fit all.

Researchers can also compare Dalarna county with other counties, especially those which are more populated like Stockholm, Malmo and Gothenburg to differentiate housing price dynamics between underpopulated and overpopulated regions.

As the results of this study are based on machine learning algorithms such as random forest, decision tree and multiple linear regression model, future researchers can use the same dataset and proceed with the "Hedonic Pricing model". Hedonic pricing model is a statistical learning model. The hedonic regression model is one of the tools that have been used extensively in analyzing the housing market in classic literature. The hedonic model estimates the house price against the house attributes (e.g. size, number of rooms, geographical area etc.) which are hypothesized to be determinants of the house price. However, this is differentiated from this work as we used machine learning algorithms which must first be trained from a set of data [14].

VIII. CONCLUSION

This work sought to establish the relationship between housing prices and the migrant population using Dalarna county as a case study. Supervised learning has been employed to build data driven models to predict the price of housing in the various Dalarna Kommuns. Though the models are reliant on the relationship between the housing prices per kommun and a number of features, the results suggest that out of the five predictor variables, the migrant population is the least determinant of the housing price.

Also, at an accuracy of 56%, it is not a model that can be recommended for use.

IX. REFERENCE

[1] N. Landberg, "The Swedish Housing Market: An empirical analysis of the real price development on the Swedish housing market.," 2016.

[2] P. De Vries. &. P. Boelhouwer, "Local house price developments and housing supply. Property management," Emerald Insight, vol. 23, no. 2, pp. 80-96, 2005.

[3] E. L. Glaeser, J. Gyourko & R. E. Saks, "Why have housing prices gone up?" American Economic Review, vol. 95, no. 2, pp. 329-333, 2005.

[4] S. Anop, "Apartment price determinants: A comparison between Sweden and Germany," Universitetsservice US-AB, Stockholm, 2015.

[5] L. Berg, "Prices on the second-hand market for Swedish family houses: correlation, causation and determinants.," European Journal of housing policy, vol. 2, no. 1, pp. 1-24, 2002.

[6] B. Keskin, "Hedonic analysis of price in the Istanbul housing market," International Journal of Strategic Property Management, vol. 12, no. 2, pp. 125-138, 2008.

[7] K. R. Corsini, "Statistical analysis of residential housing prices in an up and down real estate market: A general framework and study of Cobb County, GA," Georgia Institute of Technology, Georgia, 2009.

[8] G. James, D. Hasteie & R. Tibshirani, An introduction to statistical learning, New York: Springer, 2013.

[9] Hemnet, "Hemnet," Hemnet, [Online]. Available: https://www.hemnet.se/. [Accessed 10 04 2020].

[10] S. Sweden, "Foreign citizens by region, age in ten year groups and sex. Year 1973 - 2019," SCB, [Online]. Available: http://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__BE__BE0101__BE0101F/UtlmedbTotNK/?rxid=b83e5bbd-958a-4655-aa40-486ba2ca09a3. [Accessed 12 04 2020].

[11] D. Friedman, "Math Descriptive Statistics Article - Z-Score," Dan Friedman, 04 01 2018. [Online]. Available: https://dfrieds.com/math/z-scores.html.

[12] D. Cousineau &. S. Chartier, "Outliers detection and treatment: A review," International Journal of Psychological Research, vol. 3, no. 1, pp. 59-68, 2010.

[13] J. Clayton, N. Miller & L. Peng, "Price-volume Correlation in the Housing Market: Causality and Co-movements," The Journal of Real Estate Finance and Economics volume, vol. 40, pp. 14-40, 2008.

[14] V. Limsombunchai, "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network," NZARES Conference, pp. 1-15, 2004.