National Pension Hub

FINAL REPORT

Neighborhood Effects, Immigration and Real Estate Valuation:

A Machine Learning Approach

Authors: Erkan Yönder
Concordia University

Meriç Yücel
Istanbul Technical University

October 2021

Neighborhood Effects, Immigration and Real Estate Valuation: A Machine Learning Approach*

Erkan Yönder† *Concordia University*

Meriç Yücel‡ *Istanbul Technical University*

August 2021

Abstract

In standard econometric models, neighborhood effects are mainly ignored and loaded into location fixed effects. In this project, we merge housing data with a large set of neighborhood data and estimate house prices using machine learning models. We document that our deep learning model using neighborhood data improve prediction accuracy by 31% measured by mean absolute percentage error relative to a standard OLS model. Our findings indicate that implementing large neighborhood datasets along with machine learning models can improve the prediction accuracy of asset values. Importantly, our machine learning model reveals the positive impact of immigration on house prices.

Keywords: Real estate, valuation, neighborhood effects, immigration, machine learning, big data

*We are grateful to the Global Risk Institute and the National Pension Hub for financial support.

†John Molson School of Business, Concordia University, 1455 de Maisonneuve West, Montréal, Québec H3G 1M8, Canada. erkan.yonder@concordia.ca

‡Istanbul Technical University, Ayazaga Kampusu, Uydu Yolu, Uydu Yer Terminali Binasi (UHUZAM), Sariyer Istanbul 34469, Turkey. meric.yucel@itu.edu.tr

1 Introduction

Real estate has become an alternative asset class as being the best diversifier in pension fund portfolios (Andonov et al., 2013), sovereign wealth funds, and life insurance companies. Cvijanović (2014) states that 54% of Compustat firms reported some real estate ownership on their balance sheet. Campello and Giambona (2013) document that between 1984 to 1996, an average non-financial firm has 11.8% of its total assets invested in land and buildings, which coincides with about 33% of its tangible assets. According to Statistics Canada, the share of real estate is around 76% of the national wealth in Canada by the end of 2018. Canadian property markets are of interest by domestic and international investors increasingly in the last decade affecting the prices in the real estate markets.

Real estate valuation is complex due to its nature. First, real estate market is less liquid than stock markets or bond markets. Sales occur less frequently measured in years in the property markets. Second, real estate assets are more heterogeneous than other asset classes. Every property is unique by its characteristics, location, and neighborhood as opposed to millions of homogeneous shares of a single company being traded in seconds in the stock markets. Third, although location is fixed for any property, neighbourhoods are dynamic and change over time.

In the real estate literature, two main models have been used to value properties or create property price indices such as in S&P/Case-Shiller home price index: hedonic pricing model and repeat sales approach. Kain and Quigley (1970) are among the first to use hedonic model by controlling for physical and locational characteristics of homes. Overall, there are two main approaches to create property indices using hedonic models. In the first approach, property transactions from different years are pooled in single hedonic regression and time dummies are used to estimate the evolution of prices across time. The approach assumes the impact of hedonics on price does not change over time. In the second approach, hedonic model is estimated cross-sectionally each period using the transactions in that year. For an average property, whose characteristics are held constant, a predicted value is calculated from each period's hedonic model and the index is created by the change in the value of predicted price over time. This approach allows dynamic impact of hedonics on price by year.

On the other hand, repeat sales model is first introduced by Bailey et al. (1963) and two decades later applied by Case and Shiller (1987, 1989). The repeat sales approach uses the sale of the same property twice or more times to estimate changes in prices of same properties between two consecutive sales. The main assumption in this model is that the characteristics of properties do not change by time. The price differential of same properties is regressed on time dummies, where the time dummy gets 1 in the recent sale and gets -1 in the previous sale, and zero otherwise. The characteristics are not included in the model as they drop when taking the difference of the two consecutive sales of same property assuming that characteristics do not change by time.

Hedonic model either assumes that the coefficients of characteristics are constant over time or even if the model allows the coefficients to change, in each year, the model uses a limited number of transactions and additionally does not take neighborhood changes into consideration. On the other hand, repeat sales model assumes that the characteristics of a single property are the same over time. In addition to that characteristics do not change over time, i.e. having two bedrooms for the same property in different years, the model also assumes that characteristics are priced in the same manner across years.

Overall, these two standard models suffer from the omitted variable problem. In modeling property prices, most empirical studies ignore the quality indicators that are only observable by naked eye but use the typical hedonics of a property such as size, or the number of rooms. There are three potential approaches that can help solve the omitted variable problem.

In the first approach, the text incorporated by real estate agents is used. In a recent work, Nowak and Smith (2020) document that real estate indices created by standard repeat sales models are biased downward up to 7% during the financial crisis and upward by 20% following a crisis by using a text-based analysis on the real estate agents' posts. In another study, Liu et al. (2020) look into the text posted by real estate agents using a machine learning model (double-selection least absolute shrinkage and selection operator (LASSO)) and determine the keywords that are best predictors. Using such a model, the authors show that the premium for owner-agents that is previously documented in the literature weakens benefiting from the improvement in the prediction precision using the keywords that are best predictors. The second approach incorporates information from the images of properties. Property images can signal quality and similarity and help predict

real estate prices to overcome omitted variable problems in standard hedonic models (Lindenthal, 2020).

The third approach, which is not sufficiently incorporated into the literature, is using big data on the neighborhood characteristics that potentially change over time. Liu et al. (2020) look into the text posted by real estate agents and the authors show that although their model controls for zip code or census tract fixed effects, the model picks 5-6 neighborhood names out of 10 keywords that are best predictors of house prices. This finding indicates that using zip code or census tract fixed effects as in standard hedonic models cannot properly absorb unobserved neighborhood effects. In the literature, due to lack of availability of data, both traditional approaches in general do not take into account the dynamics of neighborhoods or how neighborhoods change over time besides using location fixed effects.

Estimations by the United Nations reflect that more than half of the world population will live in urban areas by 2050 due to immigration to cities. Internet of Things (IoT) sensor technologies are increasingly used in urban planning for issues such as air pollution, increased traffic jam, and public transportation with rapid urbanization globally. Online shopping, automated cars, industrial robots and such lifestyle changes due to technological enhancements will reshape urban living. With online shopping, the retail property markets have started to evolve. All of these technological enhancements constantly change how real estate assets are valued over time. Consistently, the COVID-19 crisis has been changing the way we live and work. In standard models, such effects due to changes in our lifestyles are assumed to be fixed or only loaded into the time or location fixed effects. Yet, it is not sufficient as evidenced by Liu et al. (2020).

In this project, we aim to develop a real estate valuation model mainly considering the dynamic impacts of changing neighborhoods such as points of interest (POI), demographics, and census information surrounding properties. We mainly aim to contribute to the literature on real estate valuation by helping solve the omitted variable problem and increase prediction accuracy. In this respect, we uniquely use big data on the dynamic neighborhoods surrounding properties. We also develop a valuation framework for academic and practical applications in the housing and commercial real estate markets.

We merge transaction data with a large set of different datasets such as historical POI data, census data, and Canadian Business Counts. To deal with such dynamic and big datasets, we

benefit from machine learning models to improve predictability power. More specifically, our aim is to improve the prediction of house prices by allowing for a larger set of neighborhood controls benefiting from machine learning models.

In general, academic research focuses on understanding relationships. Standard econometric models rely on linear relationships to be able to understand how a characteristic can affect a dependent variable of interest; the property value in our case. However, such relationships are more complex than a simple linear relationship in real life. Machine learning models can use complex (using non-linear functions) relationships other than linear effects that are used in hedonic models. Allowing for such complex effects increases the accuracy of the prediction. In a recent study, Mullainathan and Spiess (2017) improve the prediction performance of a standard hedonic model by 15-20%, on average by using machine learning models.

Using large datasets with many independent variables lowers the degrees of freedom in standard econometric models limiting the interaction effects that can be added to the model. On the other hand, machine learning models allow to use interaction effects such as in decision trees by construction (Mullainathan and Spiess, 2017). Since our aim is to use large datasets to better estimate neighborhood effects on property price prediction, machine learning models can perform better for such a task. Overall, we aim to benefit from machine learning models' capabilities to deal with interaction effects. For instance, the impact of the number of restaurants surrounding a house can be larger if there are also more office properties that are closer to the house enabling more complete work-life balance.

The application of deep learning techniques is at early stages in the academic and practical real estate literature and we also aim to contribute to this area, specifically better determining neighborhood effects on property prices. For this purpose, we use Quebec housing data. We use a transaction dataset from 2012 to 2017 and estimate house prices using a standard hedonic model and machine learning models. We then compare the performance of machine learning models to the standard hedonic model.

We develop a machine learning estimation framework that is applicable to other private asset valuations such as commercial real estate or private equity. We first use a LASSO model to determine best neighborhood predictors. Our LASSO model selects a variety of neighborhood characteristics that best predicts house prices. Overall, local income and employment are important factors affecting

house prices. Our findings also reflect the importance of immigration on house price movements as we document the neighborhood population of external immigrants and non-permanent residence increase house prices.

While previous literature mainly shows a negative association between immigration and house prices (Sa, 2015; Saiz and Wachter, 2011), Pavlov and Somerville (2020) document that an unexpected suspension of an immigration program in Canada decreases house prices. Our findings are complimentary to Pavlov and Somerville (2020) and reflect the magnitude of the impact of immigration as the LASSO model selects 6 neighborhood variables related to immigration out of 105 additional neighborhood controls. And importantly, the effect of immigration variables is net of the impact of asset characteristics, zip code fixed effects, and other neighborhood determinants.

We also show that POIs and business establishments play a role in house prices. These all indicate that the omitted variable problem can be diminished by using a big dataset on neighborhoods surrounding real estate assets. We also create a benchmark measure based on neighborhood predictors that captures house price levels to help estimation of property prices.

Once determining best neighborhood predictors, we then estimate house prices using ordinary least squares (OLS) and a wide range of machine learning models. Overall, our findings indicate that the performance of property price estimation does not significantly improve when we add neighborhood characteristics using ordinary least squares (OLS). On the other hand, in general, machine learning models improve the performance of the OLS model largely adding the neighborhood effects except the decision tree and long short-term memory (LSTM) models. Our best performing deep learning model increases out-of-sample Adjusted R-Squared of the OLS model by 16%. The mean (median) absolute percentage error go down from 23% (15%) in the OLS model to 16% (11%) in the deep learning model indicating an improvement by 31% (24%).

These findings are robust when we apply the model in different time frames. The deep learning model also consistently outperforms the OLS model across different quarters that we predict house prices. Overall, our findings demonstrate the importance of using neighborhood effects, which are largely ignored in the literature, along with machine learning models in the prediction of house prices. Using machine learning models is crucial to benefit from neighborhood effects due to the complexity of relationships.

The remainder of the paper is structured as follows. In Section 2, we discuss our datasets and how we design them. Section 3 summarizes and explains our machine learning models and estimation framework, and we document our empirical findings in Section 4. We explain practical implications for pension funds and how pension funds can benefit from our findings, and conclude in the final section.

2 Data

We use real estate data from FCIQ-Centris.ca covering single-family housing transactions occurred in Quebec from 2012 to 2017. The data include a large set of property characteristics such as sale price, building area, age of the property, number of rooms, number of bedrooms, number of bathrooms, number of garages, and whether there is a driveway. We also observe building and property type, as well as the street address of properties and the date of sale. Our aim is to create a large dataset covering neighborhood information in addition to the property characteristics used in conventional hedonic models.

Panel A of Table 1 presents the descriptive statistics for the main property characteristics. There are 246,965 transactions in our whole sample. The mean property price is $0.28 million. Average building size is 1.12 thousand sqft. The mean age of the properties in our sample is around 34 years. There are around 10 rooms and roughly 3 bedrooms in an average property. 35% of the properties have one garage and 12% of them have two garages. 90% of the properties also have a driveway.

[Table 1 about here.]

We geocode the properties in our sample using Bing Maps API. In total, we have 210,801 unique houses that are geocoded corresponding to housing transactions in our sample. We have three major sources of neighborhood data: OpenStreetMap (OSM), census data, and zip-code level business counts dataset. We match our transaction data with each of the neighborhood datasets using coordinates of each property and corresponding variables in these datasets developing Python codes. The algorithm that we develop also considers the time of the transaction and matches each transaction with the most recent observation before each transaction in addition to geographic location.

2.1 POI Data

OSM is a project to create crowd-sourced geographic database which is free to use for any type of application. All of the gathered data are accessible through the project's main website and anyone can contribute and see data about roads, railway stations, hospitals, schools and other geographical data.[1] Haklay (2010) shows that OSM database can be rather accurate with 80% of overlap in compared data comparing the roads of London. Large companies also contribute to the project as well, such as Microsoft releasing 125 million building footprints in the US that are acquired with neural networks.

OSM uses a custom Extensible Markup Language (XML) to store mapped data, i.e. buildings, streets, parks, bus stations, and additional information about these points of interests such as name, location, etc. Each item contains timestamp information storing the last modification date enabling us to observe changes in the existence of such points of interests by time.

The data format in OSM has three main elements called node, way, and relation. A node represents a single point on the map with the latitude and longitude information. Ways consists of nodes (holds references to existing nodes) and may define a street/way or a basic area, which is just like a line whose first and last node are observable. More complex shapes can be represented with relations. A relation can store any combination of references to all elements: nodes, ways and relations, which makes it possible to implement multi-polygons and draw the shapes like forests, parks, sites or complex buildings. All these three elements represent amenities such parks, restaurants, university campuses, etc.

[Figure 1 about here.]

We first determine potential amenities that can help us estimate house prices. We then extract the data on these amenities in Quebec region. In this process, we evaluate the amenities in *relations* format that are complex to geographically locate and convert them into a more structured setup. This way, we prepare geographic shape files that we can directly match with our housing dataset. The shape files are also quarterly so that we can match them with transaction data not only using the geographic location but also the time of a transaction. Figure 1 represents two example random properties with points of interests in the properties' neighborhoods.

[1] For more information, please visit http://www.openstreetmap.org.

We determine a radius surrounding an individual property as presented in Figure 1. The variables are created in a manner that for each amenity, we count the number of the amenities within a certain radius. We apply clusters at 3-km radius for each property. While such amenities are ignored at all in most similar studies, our proposed model not only includes such variables but can also incorporates the dynamics of these amenities. Figure 2 reflects an example for banks and restaurants. The figure shows how the density of banks and restaurants changes from 2015 to 2020.

[Figure 2 about here.]

2.2 Census and Canadian Business Counts Data

As our second source of neighborhood data, we obtain census and Canadian business counts data from the Statistics Canada. We download the shape files of geographic locations of divisions, subdivisions, and census tracts. Then, we obtain corresponding census data and match them with each property in our sample using geographic tools in Python. Overall, for each housing transaction, we have the corresponding census variables. We choose to use a set of variables related to unemployment rate, population, education and income level, race, and languages used in households by subdivisions. We use 2016 census data for all transactions in our sample and match them with each transaction based on location. Panel B of Table 1 presents the descriptive statistics on the census data surrounding the location of each property in our sample.

We also obtain semi-annual Canadian business counts data by divisions from Statistics Canada.[2] The data range from 2011H2 to 2017H2. We collect business count data for selected sectors, considering the overlap with the remaining neighborhood data sources for each division and each half-year. In the end, we match the data with our transaction data by time and location. Business counts data help us control the dynamics of business establishments surrounding each property in our sample. The descriptive statistics for business data is presented in Panel C of Table 1.

3 Model and Estimation Framework

Supervised, unsupervised, and reinforcement learning can be considered as the three subsets of machine learning. Our project benefits from supervised learning. In basic terms, supervised learning

[2]We download the data from http://odesi2.scholarsportal.info/documentation/CBC/cbc-en.html, which collects from Statistics Canada.

uses both features (x) and target values (y) to train the model. In testing, only features are used as inputs and the output is the predicted price (y'). The observed price values (y) and predicted value (y') are then compared to quantify the performance of the model. In our project, the features correspond to property characteristics such as amenities, and the neighborhood dataset that we create.

We first run a standard hedonic model using our data. Then, to improve the prediction accuracy of property prices, we use a variety of machine learning models. We aim to benefit from machine learning models as they allow for nonlinear relationships between independent variables and dependent variable and potentially better capture the interaction effects. We also add neighborhood data to the standard hedonic model and count on the potential of machine learning models to deal with large datasets. The machine learning models that we apply in this project are decision tree learning, random forest, gradient boosting, LSTM, and deep neural networks (DNN).[3]

In the decision tree learning method, each branch has information on observations (features' thresholds) and each leaf has the target value. For the branches, all of the features or a subset of features can be used. The training phase constructs these branches with a top-down approach (makes optimal cuts) and leaves for the given training samples. Main problems of the decision tree are that it is very sensitive to noise and it can learn too much from the train data – overfitting. Overfitting can potentially cause poor performance on unseen data.

The other approach that we use is random forest models. Random forest is an ensemble model that combines many decision trees to obtain a more generalized model. It uses the bootstrap aggregating (bagging) technique. This technique selects training samples for each decision tree randomly. These samples can also be repeated, so each set might have some redundancy. Then, each tree is trained; in case of regression, predictions can be made by averaging the predictions of those trees. For the classification problem, majority voting is used. It overcomes the sensitivity problem of decision trees because the variance of the model is decreased. While a single decision tree might be sensitive to noise, the average of them is not. Extra trees add another randomization to the random forest model. Instead of optimal cutting, cuts have been randomized. Additionally, extra trees use the whole sample data, rather than bootstrapped samples in random forest.

[3]The deep neural networks is actually a Multilayer Perception (MLP) model using neural networks. A MLP model can be considered as a subset of DNN but often are used interchangeably. For simplicity, we call it as deep learning model throughout the paper.

Gradient boosting is another ensemble learning method that uses a variation of bagging. The main idea of the method is that it uses many weak learners, usually decision trees to increase the performance. These learners are weak because it does not fit well to all of the data but can reflect some good performance on certain parts of the training. Overall, this procedure can improve the performance of the machine learning model.

LSTM is a neural network architecture which is developed by Hochreiter and Schmidhuber (1997). Unlike the general neural networks, LSTM can make feedback connections (it allows loops in the network.). LSTM architecture is mostly used in speech recognition tasks, time-series data, and anomaly detection. We also apply LSTM model to see whether we can improve prediction accuracy by benefiting from historical memory i.e. giving higher importance to more recent transactions.

The final model we apply is neural networks, McCulloch and Pitts (1943) first aim to come up with a mathematical representation of neural activity. Rosenblatt (1958) develops perceptrons, which is a basic binary classifier. Following these studies, the advancements on neural networks have been slow due to lack of computational power in the following decades. However, neural networks gain interest in the last decade. A neural network consists of neurons, which are scattered across the layers. There are three different layer types; input layer, output layer, and hidden layer. Each neuron has a weight and might be connected to the other neurons from different layers. These connections are set up from the input layer to the output layer. The hidden layer can be built by multiple layers and different operations such as non-linear activation functions (ReLU, Sigmoid), convolution, pooling, etc.

To train the data, a cost function needs to be defined. After each learning step, the cost function is evaluated. The cost should be diminished periodically to achieve learning and when an optimal cost value is achieved, the model can predict y using x. To improve the performance of the model, learning rate, number of layers, number of neurons in each layer, which activation function to use are needed to be tuned.

We present a generalised block diagram of our estimation framework in Figure 3. We merge our standard transaction data with the neighborhood data that we create. Neighborhood data consist of OSM's POI data, census, and business establishment data. The final merged data consist of 105 neighborhood variables, which decreases the degrees of freedom in standard hedonic models and has very high dimensionality.

[Figure 3 about here.]

High dimensionality may cause several problems in machine learning models such as overfitting, reduced prediction accuracy, insufficient training, time and memory complexity, etc. To overcome such problems, we apply dimensionality reduction to the input data set. Dimensionality reduction can be applied in two different ways: the reduction by feature selection or component-based reduction methods such as principle component analysis, projection-based ISOMAP, t-SNE,UMAP, etc.

We use the feature selection method to reduce dimensionality to keep real values in input data set rather than projected variables. Our feature selection process also enables us to explain what neighborhood factors best predict house prices. For this purpose, we use a LASSO model where we include neighborhood data in addition to the standard hedonic model. LASSO model aims to minimize the absolute value of the coefficients and, therefore produces many coefficients that are zero, which makes it favourable for feature selection (Tibshirani, 1996). Overall, the LASSO model forces the coefficients to become zero and determines the best predictors that remain non-zero. This way, we choose best predictors as features to be included in the machine learning models.

Once we determine the best neighborhood predictors, we then divide the data into train and test data groups. In our main setting, we exclude the transactions that occur in the very final quarter of our sample (2017Q4) and use that quarter as our test group and all the transactions in the previous quarters as the train group from 2012Q1 to 2017Q3. We normalize the train data with standard scalars. We standardize features by subtracting mean and scaling to unit variance. As explained above, we perfectly separate the train and test data. Accordingly, the mean and standard deviation of the train data are used to normalize the test data, as well.

We also do a recursive-type analyses where we start our train data from 2012Q1 to the end of 2013Q4 and predict property prices in 2014Q1 using the standard hedonic model and our machine learning models. Then, we repeat the process for the transactions between 2012Q1 and 2014Q1 as the train group and 2014Q2 as the test group. We roll our sample this way to have predictions for each quarter till the end of our sample. Finally, we compare model performances.

4 Results

4.1 Best Neighborhood Predictors and Immigration

The LASSO model results reflect important findings. In a census tract, higher population density, a higher number of dwellings, and the number of high-rise apartments increase house prices. More higher-income households in a location also increase house prices. Interestingly, the share of population aged 65 and over also positively affects house prices. Higher local unemployment rate and the share of subsidized housing lower house prices. On the other hand, higher shelter costs are also positively associated.

[Table 2 about here.]

Regarding immigration and minority population, higher share of minority official language, the Arab and Southeast Asian population, external immigrants, and non-permanent residents are associated positively with house prices. Conversely, the population of interprovincial immigrants are negatively associated with house prices. This indicates the role of immigration on the rise of house prices in Canada.

Our machine learning model clearly captures the immigration impact and chooses 6 neighborhood variables related to immigration out of 105 neighborhood characteristics. The positive impact of immigration on house prices that is reflected by our model also contributes to the real estate literature on immigration. While our findings show an opposite impact of immigration compared to Saiz and Wachter (2011) and Sa (2015) but are complimentary to the recent work by Pavlov and Somerville (2020). Importantly, our findings on immigration are after controlling for a large set of neighborhood characteristics, which is unique to the literature.

POIs and business establishments also reflect significant impacts on house prices. Public transport and bicycle options are positively associated with price. On the other hand, bicycle parking negatively affects house prices potentially reflecting a nonlinear relationship. Electric car charging stations are also positively associated with house prices. The number of restaurants, cafes, kindergartens, colleges, book stores, and fitness centres are all selected by the LASSO model.

Among business establishments, management of companies and enterprises, miscellaneous store retailers, and warehousing and storage are positively associated. Alcoholic drinking places and

department stores have a negative relationship with house prices. Overall, our findings using the LASSO model reflect the impacts of the neighborhood variables that are omitted mostly in standard model, which in turn potentially affects prediction accuracy and raises doubts over the omitted variable problem in those models.

[Figure 4 about here.]

We also create a benchmark measure using best neighborhood predictors. Figure 4 reflects the net contribution of the best neighborhood predictors to house prices and compares it with the heat map of actual prices. As seen from the figure, the heat map created based on the neighborhood predictors is in line with the heat map of actual prices indicating that a benchmark created based on house prices help improve prediction accuracy of house prices. Our benchmark measure is potentially helpful to identify hot markets based on neighborhood characteristics.

4.2 Baseline Results

We start our analysis by first estimating house prices using standard hedonic models.[4] Table 3 presents the results. In column 1, we run house prices on property characteristics without using zip code fixed effects. We add zip code fixed effects in column 2. The in-sample Adjusted R-squared increases from 49% to 79% when we add zip code fixed effects. In column 3, we add neighborhood characteristics that are selected by our LASSO model without controlling for zip code fixed effects. Adjusted R-squared is close to the regression in column 2, that is 76%. Best neighborhood predictors reflect a similar performance to zip code fixed effects, without controlling for the zip code itself.

We also control for our neighborhood benchmark indicator in column 4. Adjusted R-squared increases from 49% (column 1) to 70%. The advantage of our benchmark measure is that it increases degrees of freedom in a regression model by using the benchmark (only one indicator) instead of using a high number of location fixed effects or neighborhood variables and can obtain similar predictability power. This indicates that our neighborhood benchmark is a strong predictor of house prices, which is also statistically significant at 1% level.

[4] In our machine learning models and hedonic model that we compare the performance with of those models, we use the price, itself as the dependent variable. Our unreported analysis, using nominal price gives better prediction for out-of-sample analysis than the logarithm of price. On the other hand, in Table 3, the dependent variable is the logarithm of price to report better explainable coefficients in the table.

In column 5 and 6, we control for the neighborhood characteristics and our neighborhood benchmark, respectively, in addition to zip code fixed effects. Adjusted R-squared increases by around 1% in both specifications. Importantly, our neighborhood benchmark is still significant at 1% level after controlling for zip code fixed effects. This indicates that zip code fixed effects do not sufficiently control for neighborhood effects and the neighborhood benchmark's impact is still statistically significant net of zip code fixed effects.

In general, the coefficients of property characteristics have expected effects. 1% increase in building area increases house price by 1.8%. Older properties priced worse while there is a nonlinear relationship between age and house price. House prices increase by property characteristics such as the number of rooms, bedrooms, and bathrooms. Having a second garage increases house prices by 21% while having only one garage increases by 8%.

[Table 3 about here.]

4.3 Relative Performance of Machine Learning Models

We then turn our attention to the estimation of machine learning models. Table 4 presents the comparison of performance of machine learning models relative to the standard hedonic model. In all models including the OLS model, the sample period for the train data is from 2012Q1 to 2017Q3 and the sample period for the test data is 2017Q4. We use the LASSO model with neighborhood data to determine best predictors and to select best features in the machine learning model. We also use the OLS model with all features as the base model to report performance improvement that different machine learning models make.

[Table 4 about here.]

Adjusted R-squared of the OLS model is 67%. The MAPE and the MdAPE of the OLS model are 23% and 15%, respectively. This indicates that overall, the OLS model makes an error by 15% on an average transaction, if we use MdAPE. In other words, if the true price of a property is $1 million, OLS on average predicts a price of $850,000 or $1,15 millions.

Overall, machine learning models perform much better than the standard hedonic model using the neighborhood data. Overall, Adjusted R-Squared increases by at least 8.4% up to 16% of that

of the OLS model except the decision tree and LSTM models. The improvement in MAPE and MdAPE is much larger up to 31% and 24%, respectively. The overall improvement in our paper is larger than the 15-20% improvement in Mullainathan and Spiess (2017). The best performing model is our deep learning model. The median of the absolute percentage error goes down to nearly 11% without any sample selection in the data. This indicates that if the true price of a property is $1 million, deep learning model on average predicts a price of $890,000 or $1,11 millions.

The random forest and extra trees models also perform well with around 22% to 25% improvement in MAPE. While the gradient boosting model improves the performance gradually compared to other machine learning models, the only models that perform worse than the OLS model is the decision tree model and the LSTM. The Adjusted R-Squared decreases to 56% relative to 67% of the OLS model. MAPE and MdAPE also increases to 26% and 17%, respectively, in the decision tree model.

We also analyze the performance of our model by time to test whether our findings are time- or sample-specific. We start the train data using the period between 2012Q1 and 2013Q4 and predict the prices in 2014Q1 using as the test sample. Then, we rerun the model by increasing the end of the train sample by one quarter and moving the test sample by also one quarter. Figure 5 presents rolling quarterly MAPEs of the OLS model and the deep learning model. In all quarters, the deep learning model consistently beats the OLS model by 30-40% of the MAPE of the OLS model.

[Figure 5 about here.]

Overall, our analysis reflects the importance of using neighborhood data and how better machine learning models deal with such big datasets and improve the prediction accuracy. Our deep learning model consistently reflects better performance compared to the standard hedonic model using quintiles or different time periods to predict house prices. Our findings reflect the magnitude of the omitted problem and how using neighborhood data can improve prediction accuracy.

5 Practical Implications for Pension Funds and Concluding Remarks

Most empirical works to value assets such as private equity, real estate, or even the stocks of publicly listed firms do not control for neighborhood effects. In general, in standard empirical setups, dummy variables for each location (location fixed effects) are used to capture any neighborhood effects. For

instance, counting state mentions in 10-Ks, García and Norli (2012) and show that local firms have better returns than geographically diversed firms.

While even the counts of location names can help improve our understanding, machine learning models can help deal with large sets of neighborhood variables and implement into the models. Liu et al. (2020) go beyond word counts and use a machine learning approach to determine keywords that are best predictors of house prices. Importantly, the authors find that using location fixed effects does not sufficiently capture neighborhood effects in house prices. Their findings strongly indicate that we need to dig into the omitted variable problem for neighborhood effects. Omitting neighborhood indicators can affect prediction accuracy and lead to misvaluations and wrong inferences. This problem is not limited to the housing problem but it also applies to the valuation of other asset classes such as commercial real estate, private equity, and stocks.

[Figure 6 about here.]

We develop an analytical framework to benefit from neighborhood data in the estimation of asset values. Figure 6 generalizes our framework for different asset classes. Specifically, in this project, we use Quebec housing data to document how neighborhood indicators can improve prediction accuracy along with machine learning model applications. We merge a standard housing transaction data with neighborhood datasets such as OSM, census, and Canadian Business Counts, and compare the performance of a standard hedonic model with different machine learning approaches.

Overall, we find that the inclusion of neighborhood datasets along with using machine learning models largely improve the performance of the OLS estimation of house prices. The deep learning model that we develop increases the Adjusted R-Squared of the OLS model by 11.5%. The median of the absolute percentage error goes down to around 11% with an improvement by 24%. Importantly, our analysis reveals the positive impact of immigration on house prices.

Our findings have important implications. Using large sets of neighborhood indicators improves model performance economically significantly if machine learning models are applied. Based on our machine learning model, we develop a neighborhood benchmark score, which can help investors better assess their assets and investments using our neighborhood benchmark. Our findings are not limited to the housing markets but they are also potentially applicable to other asset classes such as commercial real estate, private equity or stocks of publicly listed firms. Based on similar machine

learning models, dynamic neighborhood benchmarks can be developed and prediction accuracy can be improved by adding neighborhood benchmarks to standard OLS models.

References

Andonov, Aleksandar, Nils Kok, and Piet Eichholtz, 2013, A Global Perspective on Pension Fund Investments in Real Estate, *Journal of Portfolio Management* 39, 32–42.

Bailey, Martin J., Richard F. Muth, and Hugh O. Nourse, 1963, A Regression Method for Real Estate Price Index Construction, *Journal of the American Statistical Association* 58, 933–942.

Campello, Murillo, and Erasmo Giambona, 2013, Real Assets and Capital Structure, *Journal of Financial and Quantitative Analysis* 48, 1333–1370.

Case, Karl, and Robert Shiller, 1987, Prices of Single Family Homes Since 1970: New Indexes for Four Cities, Cowles Foundation Discussion Papers 851, Cowles Foundation for Research in Economics, Yale University.

Case, Karl E., and Robert J. Shiller, 1989, The Efficiency of the Market for Single-Family Homes, *American Economic Review* 79, 125–137.

Cvijanović, Dragana, 2014, Real Estate Prices and Firm Capital Structure, *Review of Financial Studies* 27, 2690–2735.

García, Diego, and Oyvind Norli, 2012, Geographic dispersion and stock returns, *Journal of Financial Economics* 106, 547–565.

Haklay, Mordechai, 2010, How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets, *Environment and Planning B: Planning and Design* 37, 682–703.

Hochreiter, Sepp, and Jürgen Schmidhuber, 1997, Long Short-term Memory, *Neural Computation* 9, 1735–80.

Kain, John F., and John M. Quigley, 1970, Measuring the Value of Housing Quality, *Journal of the American Statistical Association* 65, 532–548.

Lindenthal, Thies, 2020, Beauty in the eye of the home-owner: Aesthetic zoning and residential property values, *Real Estate Economics* 48, 530–555.

Liu, Crocker H, Adam D Nowak, and Patrick S Smith, 2020, Asymmetric or Incomplete Information about Asset Values?, *Review of Financial Studies* 33, 2898–2936.

McCulloch, Warren S., and Walter Pitts, 1943, A Logical Calculus of the Ideas Immanent in Nervous Activity, *Bulletin of Mathematical Biophysics* 5, 115–133.

Mullainathan, Sendhil, and Jann Spiess, 2017, Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives* 31, 87–106.

Nowak, Adam D., and Patrick S. Smith, 2020, Quality-adjusted house price indexes, *American Economic Review: Insights* 2, 339–356.

Pavlov, Andrey, and Tsur Somerville, 2020, Immigration, capital flows and housing prices, *Real Estate Economics* 48, 915–949.

Rosenblatt, Frank, 1958, The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain., *Psychological review* 65 6, 386–408.

Sa, Filipa, 2015, Immigration and house prices in the uk, *Economic Journal* 125, 1393–1424.

Saiz, Albert, and Susan Wachter, 2011, Immigration and the neighborhood, *American Economic Journal: Economic Policy* 3, 169–188.

Tibshirani, Robert, 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 267–288.

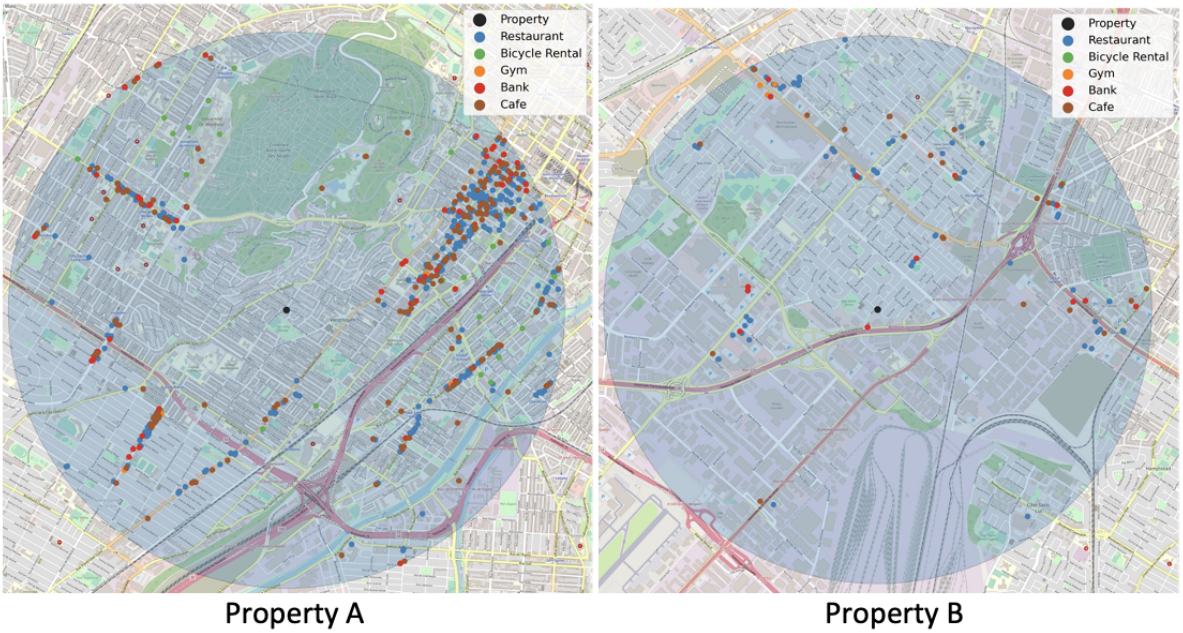

Figure 1. Amenities in the Clusters surrounding Individual Properties

The figure presents the amenities such as restaurants, bicycle rentals, gyms, banks, and cafes surrounding two representative properties.

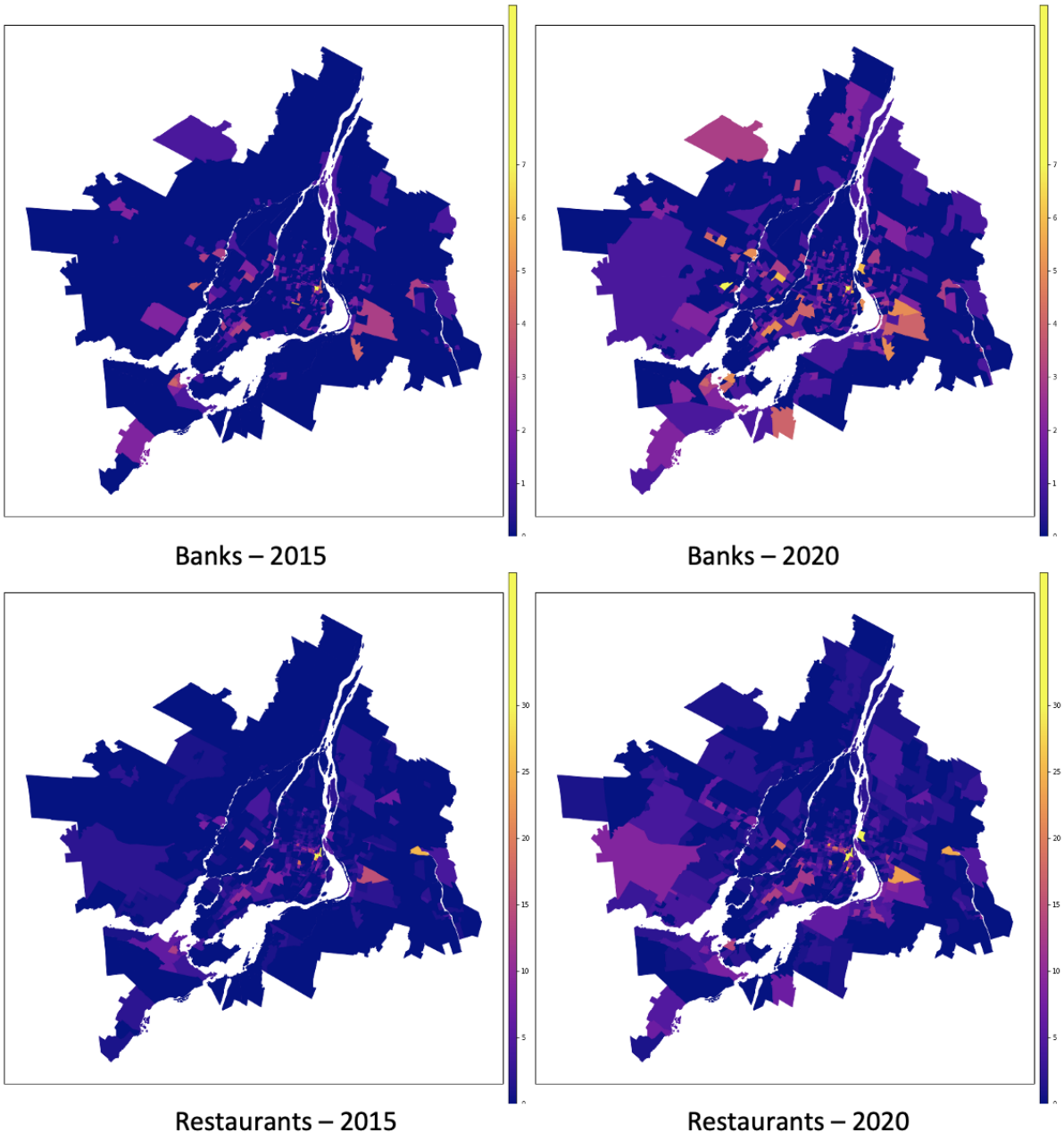

Figure 2. Changes in the Density of Amenities by Time in Montreal Area

The figure shows the change in the density of banks and restaurants in Montreal Area from 2015 to 2020.

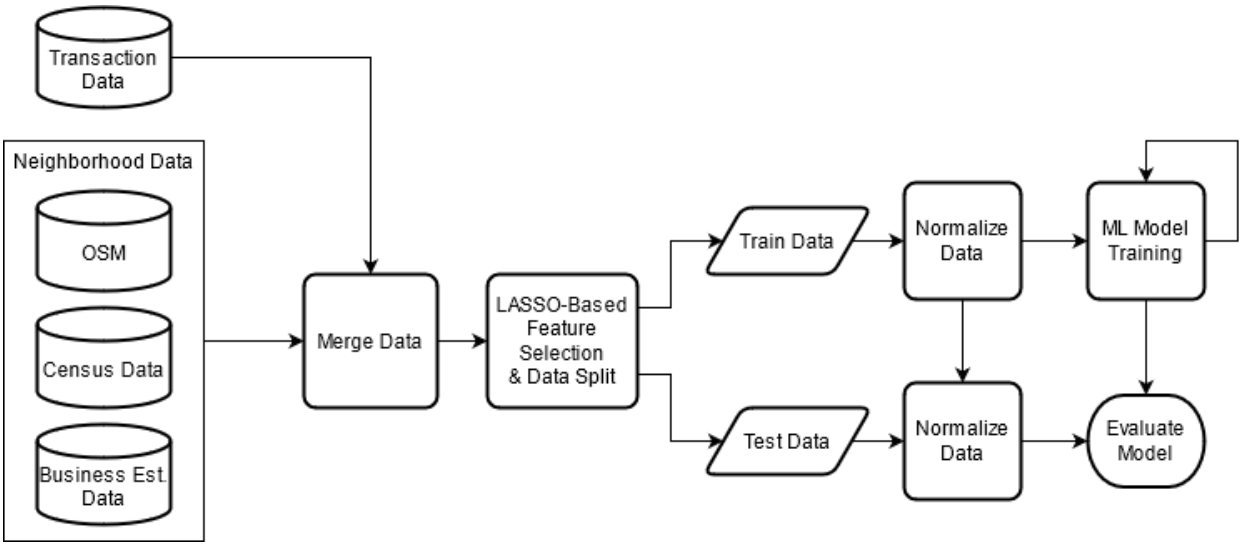

Figure 3. Model Diagram

The figure presents a standard model diagram for all machine learning models applied..

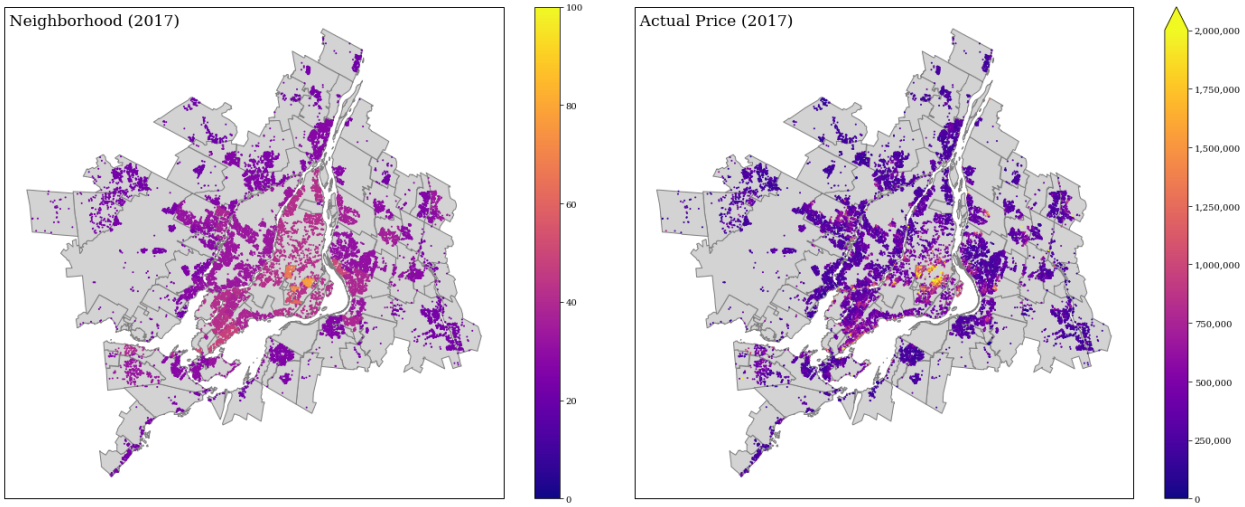

Figure 4. Contribution of Best Neighborhood Predictors to House Price Prediction

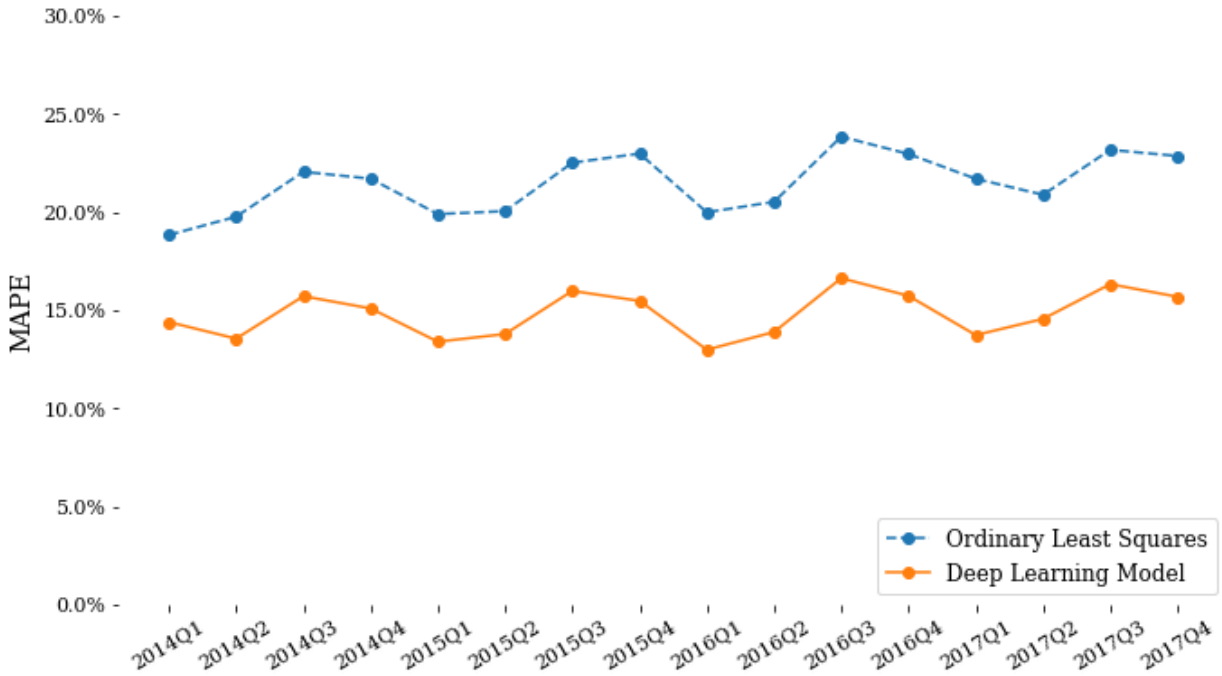

Figure 5. Mean Absolute Percentage Error by Rolling Quarterly Predictions

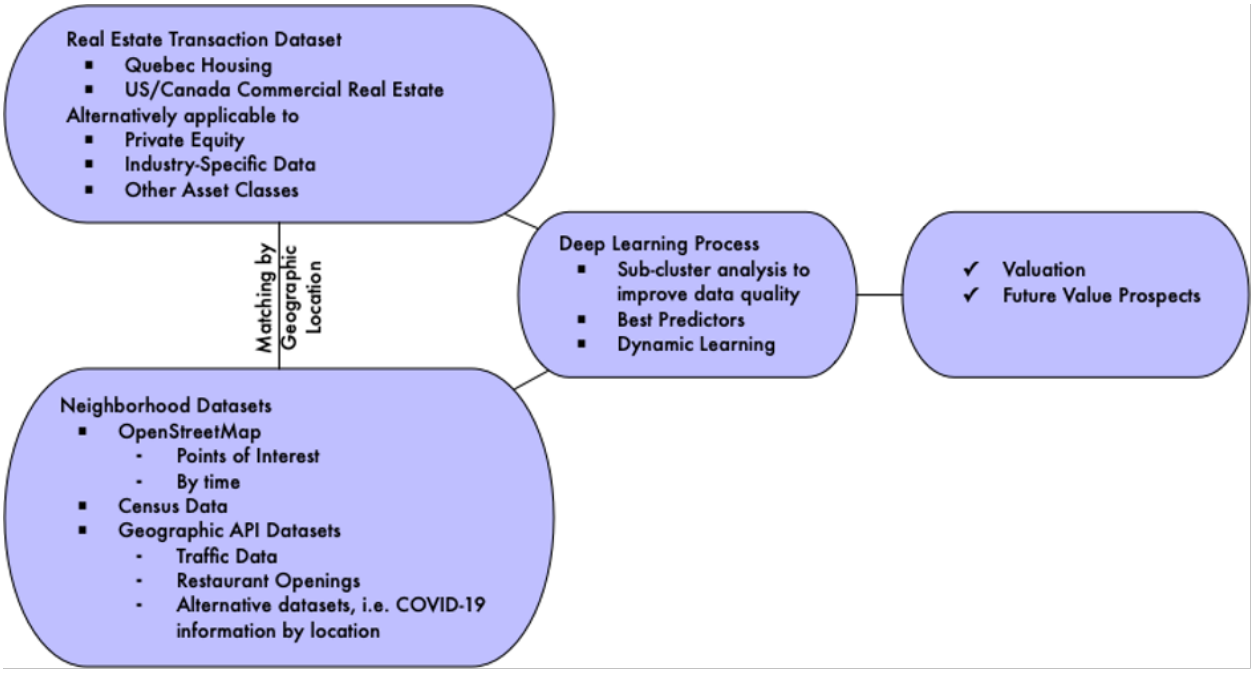

Figure 6. Analytical Framework for Neighborhood Effects

The figure presents the analytical framework for neighborhood effects.

Table 1. Descriptive Statistics

	Mean	SD	Min	Max	N
	Panel A - Property Characteristics				
Price ($ million)	0.28	0.19	0.05	13.25	246,902
Building Area (thousand sqft)	1.12	68.79	0.00	31,448.54	246,902
Property Age	33.78	24.87	0.00	200.00	246,902
# of Rooms	9.93	2.74	1.00	38.00	246,902
# of Bedrooms	2.64	0.84	1.00	10.00	246,902
# of Bedrooms Plus	0.77	0.83	0.00	9.00	246,902
# of Bathrooms	1.53	0.61	0.00	5.00	246,902
# of Bathrooms Plus	0.48	0.55	0.00	5.00	246,902
1 Garage	0.35	0.48	0.00	1.00	246,902
2 Garages	0.12	0.32	0.00	1.00	246,902
2+ Garages	0.02	0.14	0.00	1.00	246,902
Driveway	0.90	0.30	0.00	1.00	246,902
	Panel B - Selected Census Data				
Income as a Share of Population					
$50,000 to $59,999	0.07	0.01	0.02	0.11	246,902
$60,000 to $69,999	0.05	0.01	0.00	0.12	246,902
$70,000 to $79,999	0.04	0.01	0.00	0.08	246,902
$80,000 to $89,999	0.03	0.01	0.00	0.06	246,902
$90,000 to $99,999	0.02	0.01	0.00	0.05	246,902
$100,000 and over	0.05	0.03	0.00	0.22	246,902
Owner Private Households	0.69	0.15	0.37	1.08	246,902
Renter Private Households	0.31	0.15	0.00	0.63	246,902
Median Government Transfers ($)	7,199.31	1,440.2	3,000.00	13,992.00	246,902
Unemployment Rate (%)	6.63	2.21	0.00	38.60	246,902
Non-permanent Residents	0.01	0.01	0.00	0.06	246,902
Official Language Minority (%)	11.60	15.37	0.00	93.80	246,902
Interprovincial Migrants	0.01	0.02	0.00	0.24	246,902
External Migrants	0.04	0.05	0.00	0.25	246,902
Minority Population in Private Households					
Chinese	0.01	0.02	0.00	0.14	246,902
Arab	0.02	0.02	0.00	0.08	246,902
Southeast Asian	0.01	0.01	0.00	0.04	246,902
Main Mode of Commuting					
Car, Truck, Van - as a Driver	0.37	0.06	0.10	0.52	246,902
Public Transit	0.04	0.05	0.00	0.16	246,902
Walked	0.02	0.01	0.00	0.11	246,902
Bicycle	0.00	0.00	0.00	0.03	246,902
	Panel C - Business Counts Data				
Finance and Insurance	1,375.62	2,726.00	4.00	12,468.00	246,902
Management of Companies and Enterprises	678.13	1,554.76	0.00	6,461.00	246,902
Non-Store Retailers	114.52	191.31	0.00	875.00	246,902
Hospitals	8.12	18.39	0.00	89.00	246,902
Miscellaneous Store Retailers	187.67	343.06	0.00	1,557.00	246,902
Department Stores	3.62	6.51	0.00	29.00	246,902
Warehousing and Storage	27.40	55.63	0.00	260.00	246,902
Drinking Places (Alcoholic Beverages)	66.92	137.63	0.00	633.00	246,902

The table presents the descriptive statistics for the property characteristic, the selected census data, and business establishment data.

Table 2. Best Neighborhood Predictors Selected by the LASSO Model

	LASSO Sign
Census Data	
Dwellings	(+)
Population Density per Square Kilometer	(+)
Apartment in a Building that Has 5 or more Storeys	(+)
55 to 59 Years	(−)
65 Years and over	(+)
Median Government Transfers ($)	(−)
$50,000 to $59,999	(+)
$60,000 to $69,999	(+)
$70,000 to $79,999	(+)
$80,000 to $89,999	(+)
$100,000 and over	(+)
$150,000 and over	(+)
Median Income of One-Person Households ($)	(+)
Owner	(−)
% of Owner Households Spending 30% or more of Its Income on Shelter Costs	(+)
% of Tenant Households in Subsidized Housing	(−)
Unemployment Rate	(−)
Official Language Minority (%)	(+)
Non-Permanent Residents	(+)
Arab	(+)
Southeast Asian	(+)
Interprovincial Migrants	(−)
External Migrants	(+)
Public Transit	(+)
Bicycle	(+)
Business Counts Data	
Management of Companies and Enterprises	(+)
Miscellaneous Store Retailers	(+)
Warehousing and Storage	(+)
Department Stores	(−)
Drinking Places (Alcoholic Beverages)	(−)
OpenStreetMap (Counts in 3-km Radius)	
Charging Station	(+)
Bicycle Parking	(−)
Public Transport	(+)
Restaurant	(+)
Cafe	(+)
Kindergarten	(+)
College	(+)
Bank	(+)
Books	(−)
Fitness	(−)

The table presents best predictors determined by the LASSO model and the direction of the impact on house prices.

Table 3. Standard Hedonic Model Estimations

	(1)	(2)	(3)	(4)	(5)	(6)
ln(Building Area)	0.031***	0.018***	0.023***	0.028***	0.018***	0.018***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Property Age	-0.005***	-0.011***	-0.01***	-0.01***	-0.011***	-0.011***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Squared Property Age	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
# of Rooms	0.006***	0.026***	0.024***	0.022***	0.026***	0.026***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
# of Bedrooms	0.073***	0.018***	0.022***	0.033***	0.019***	0.018***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
# of Bedrooms Plus	0.020***	-0.001	0.003***	0.011***	0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
# of Bathrooms	0.306***	0.140***	0.148***	0.162***	0.131***	0.135***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
# of Bathrooms Plus	0.210***	0.087***	0.089***	0.105***	0.080***	0.083***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
1 Garage	0.117***	0.080***	0.083***	0.084***	0.082***	0.081***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
2 Garages	0.224***	0.207***	0.212***	0.187***	0.207***	0.206***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
2+ Garages	0.147***	0.204***	0.205***	0.167***	0.206***	0.204***
	(0.005)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Driveway	-0.011***	0.023***	0.008***	0.026***	0.021***	0.023***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Neighborhood Score	—	—	—	0.040***	—	0.021***
				(0.000)		(0.000)
Constant	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Property-Type FE	Yes	Yes	Yes	Yes	Yes	Yes
Building-Type FE	Yes	Yes	Yes	Yes	Yes	Yes
Postal Code Effects	—	Yes	—	—	Yes	Yes
Neighborhood Effects	—	—	Yes	—	Yes	—
Adj. R-squared (In-sample)	0.485	0.788	0.756	0.702	0.802	0.794

The table presents the OLS estimation of the standard hedonic model. The dependent variable is the logarithm of house prices. Fixed effects are included as indicated. Robust standard errors are reported in parentheses. Statistical significance is indicated as follows: * p <0.1; ** p <0.05; *** p <0.01.

Table 4. Relative Performance of Machine Learning Models over OLS Model

	Out-of-Sample Accuracy			Relative Improvement over OLS		
	Adj. R-Squared	MAPE	MdAPE	Adj. R-Squared	MAPE	MdAPE
OLS (All Features)	0.667	22.846	14.940	—	—	—
Decision Tree	0.563	25.747	16.535	-15.59%	-12.70%	-10.68%
Random Forest	0.768	17.750	11.375	15.14%	22.31%	23.86%
Gradient Boosting	0.723	20.955	14.144	8.40%	8.28%	5.33%
Extra Trees	0.775	17.036	11.063	16.19%	25.43%	25.95%
LSTM	0.656	16.936	12.282	-1.65%	25.87%	17.79%
Deep Learning Model	0.744	15.676	11.283	11.54%	31.39%	24.48%

The table presents the relative performance of machine learning models compared to OLS estimation. The performance indicators are Adjusted R-Squared, mean absolute percentage error, and median absolute percentage error.