Final Report

GTA HOUSE PRICE PREDICTION

Team 1

Table of Contents

Executive Summary

The project is designed to predict the price of a non-commercial house in the GTA area by building different predictive models.

The non-commercial housing market is booming in Canada, especially in metros like the GTA. There is an increasing demand to know the appropriate price of the real estate for both buyers and sellers. Given the large amount of historical transactions and abundant housing data, the project is trying to explore the relationship between the features of a house and its final sold price. However, other factors such as economic trends and the state of public infrastructure can also influence house price.

We will use a housing-profile dataset from the Toronto Regional Real Estate Board (MLS) and Toronto Open Data to build four models to predict house prices across GTA area. We will further explore the Google Maps API to integrate locations of infrastructures (schools, hospitals, major highways etc.) that may influence house prices in the area.

The models that are evaluated include LU Decomposition, Linear Regression, Decision Trees and Neural Networks. After evaluating each model, we apply a simple average ensemble method to explore if the prediction can be improved. From the evaluation, we found that all the models have returned a high R-Square value. The reason behind the high value is due to the influence of the asking price which has been included in the dataset. This feature has served as a baseline to determine the house price. In real life, one may not be able to set an asking price. The second part of the project focuses on the datasets without the asking price feature, and we apply the same models. The winner is the neural network model.

Business Problem Statement

In a capitalistic society, the house prices are determined by the free market. True value of a house is only what buyers are *willing* to pay for it. Moreover, in Toronto, a house can only be put on the market by a licensed real estate agent to facilitate the transaction, which typically costs five percent of the price of the house. This means a buyer can't accurately predict the true value of their property without serious upfront commitments.

When listing a house, the 'time it has been on the market' is shown on all listings in Toronto. It is common to associate longer wait times with the idea that "there may be something wrong" with a property. Since most buyers are first time buyers or want to invest in a house in which they can live in for at least a medium term, they tend to be very cautious of their purchases.

When someone is selling a real estate property, their objective is to optimize the asking price such that it minimizes the time on market, while getting the most money for it. This is especially important when you look at the scale of this problem. Just in the City of Toronto, there were more than a hundred thousand houses sold in 2019 alone, which is a fairly high transaction volume.

However, there are a few methods currently available that help sellers get a rough estimate for the worth of their properties and help buyers plan budgets. These include:

i. **House Price Index**: Various House Price Indexes are published by financial intuitions across Canada, e.g. by the Bank of Canada, every quarter. However, such indexes do not have the flexibility to include house specific information that can have a direct impact on the house price. For example, they don't consider if the energy source is gas or electricity, or if a major renovation work was completed recently – both of which can significantly influence the final price.

ii. **Find comparable properties:** An alternative is to find comparable properties to estimate a house price. For example, if House A and B are similar, and B was sold for $X, with a certain confidence level, it can be deduced that property A might also be sold for the same amount. The challenge is getting hold of such data, especially in the age where most governments and private sector organizations are very protective of their data. In fact, finding the right dataset to build our model was perhaps the biggest challenge that we encountered. This could be even more challenging for buyers and sellers who may not be as data savvy.

iii. **Hiring Professional Appraisers:** A third alternative is to hire a professional appraiser to evaluate property prices. In Toronto, professional appraisers are the same as real estate agents, who rely on their experience and knowledge to assess house values. However, like all humans, they are also susceptible to their own biases and limited by their knowledge of the area. On top, given the competitive housing market over the last few years, most real estate agents would over promise prices just to acquire clients.

After conducting interviews with Master of Real Estate Management students at Schulich School of Business, our team felt that there was a clear need for a data driven, Artificial Intelligence model. This model can work alongside humans to predict, with a high accuracy, the selling prices for real estate properties across Toronto.

A tool like this will be useful for sellers to better understand the value of their property without having to go through an extensive exercise of hiring agents, placing adds, and negotiating prices, and risk underselling. It will be useful for buyers to plan budgets, especially when there are pricing wars for properties. Finally, it will be useful for real estate agents who can enhance their credibility by demonstrating use of cutting-edge technologies to offer a better customer service.

The next section discussed the methodology that was used to collect and pre-process data, select, train and test models and overall performance evaluation.

Methodology

The methodology of the project can be broken down into three phases:

1. Data collection and preparation
2. Model selection, build, training and validation
3. Model evaluation

Data collection and preparation

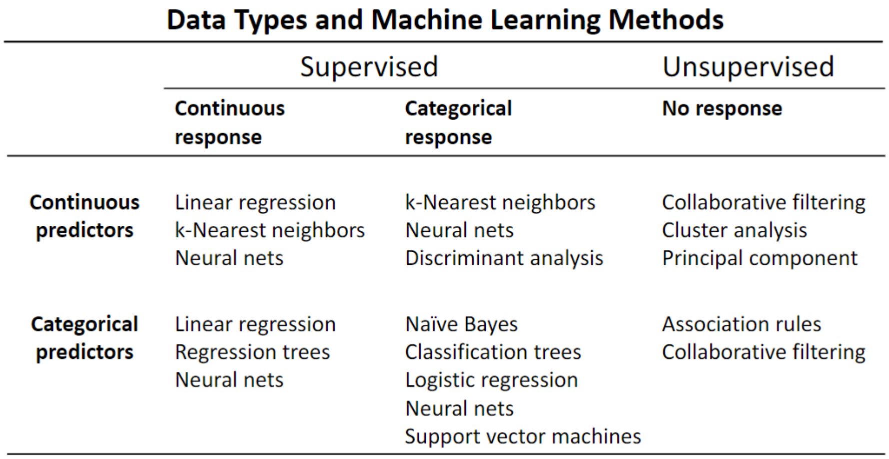
The dataset which includes approximately 10,000 transactions, comes from two resources:

1. Toronto Real Estate MLS (Multiple Listing Service): it is a portal that stores historical transactions for every property.
2. Toronto Open data: from this resource, we were able to find locations of schools and apply web-scraping technique to calculate the geographic distance for each data point.

After data extraction, we combined these two data sources. Our final dataset includes 12 features and one output, which is final sold price.

Model selection, build, training and validation

We leveraged the following chart to narrow down the choices of models we could apply to our real-estate price prediction problem. Since our dataset was labelled, it fell under Supervised learning. In addition, given the continuous response (i.e. house prices), we were able to focus on the first column of the chart.



Data Types and Machine Learning Methods

	Supervised		Unsupervised
	Continuous response	**Categorical response**	**No response**
Continuous predictors	Linear regression k-Nearest neighbors Neural nets	k-Nearest neighbors Neural nets Discriminant analysis	Collaborative filtering Cluster analysis Principal component
Categorical predictors	Linear regression Regression trees Neural nets	Naïve Bayes Classification trees Logistic regression Neural nets Support vector machines	Association rules Collaborative filtering

Overall, we have selected four models. These include Linear Regression Model, Neural Net Model, and Decision Tree Model. In one of our other courses this semester, we had also been taught a computationally inexpensive model known as LU Decomposition. This seemed like an excellent opportunity to put it to test, therefore LU Decomposition was also selected.

Linear Regression and LU Decomposition are similar in the way that they both leverage a line of best fit; and return a linear equation that can predict the final price. The Decision Tree Model can produce a tree-like graph to show how the model arrives at the final price. The Neural Networks, even though they work like a black box, have been receiving widespread attention due to their high accuracy.

After selecting our models, the team built, trained and tested them on our test dataset. We also applied the ensemble method to see if the collection of all these models will improve the overall performance. For this project, simple average ensemble method has been used.

Model evaluation

We used a regression summary and R square score to evaluate each model based on the validation dataset. As part of the regression summary statistics, we calculated the Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). We decided to use R square score as our main performance indictor because it is more straight-forward and intuitive to understand in our application and is widely accepted as a measure to evaluate and compare model performance.

Data Collection & Preparation

Data Sources & Integration

We used two data sources to construct our data frame consisting of a mix of an open and a closed dataset. These include the closed Toronto Regional Real Estate Board/MLS dataset provided to us by a licensed-realtor colleague. This dataset had property details and prices for over 15,000+ properties. Since only ~10,000+ properties had been sold, i.e. had an actual sold price, only these properties were used to train and develop our models. We also used an open data-source from the Toronto Open Data Portal which has the geolocation data of over 500 schools. Some of these schools were omitted as they were outside the housing regions under consideration for our project. For each school, the dataset provided geodesic distances of schools from the properties. The rationale to include the school distance feature is that house prices increase if the property is located in close proximity to certain landmarks, including schools, hospitals, shopping malls, grocery stores, highways and parks. Amongst these, schools are at a premium as most people who can afford to buy a house in Toronto are at an age where they would already, or in the future, have school-going children. The remaining landmarks can be included in future iterations of our MVP.

Independent Variables/Features

The MLS dataset holds all the critical profile data on a house, which were used as the predictors / independent variables. These included house address, final price (actual price the house sold for), list price, number of bedrooms, bathrooms, square footage, dens, parking, and mean neighbourhood income. The Toronto Open Data Portal provided the longitudinal and latitudinal features for schools.

Data quality and Relevance

Overall, the MLS data was of reasonably high quality but required cleaning nonetheless- because some features had missing values, some had descriptive fields (e.g. details of recent renovations), and some fields simply did not hold much value as predictors (e.g. advertisement listing numbers). Such fields were discarded for ease in pre-processing the data. As part of a future iteration, if any Natural Language Processing techniques are employed to improve the performance of the model, the descriptive fields can be included to train such models.

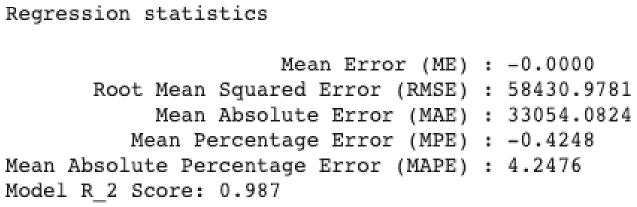
In addition, another concern was detecting any seasonality in our data. Its common knowledge that real estate prices tend to increase over time, and therefore it is expected that there would be seasonality in our dataset. Unfortunately, our MLS dataset did not include a date of purchase / sale as a feature, which meant that despite our best efforts, we could not address the issue of seasonality. Upon further investigation it was found that, first, such data is only available for purchase, and second, that all transactions in our dataset spanned over a six-month time window. Given the small-time horizon, and our budget limitations, it was assumed that any seasonality in our dataset is small enough to be ignored. Perhaps if we were to improve one thing in our model, we will purchase datasets that include a time element to retrain our models with seasonality and trend adjusted data.

Model Analysis & Results

Before beginning with model analysis and takeaways, we conducted an exhaustive search to determine the best predictors to use as inputs. Based on our results, we determined that using all features produced the best results.

LU Decomposition

The LU decomposition method involves the factorization of a given square matrix into two triangular matrices, one upper triangular matrix and one lower triangular matrix, such that the product of these two matrices gives the original matrix. It was introduced by Alan Turing in 1948. A lot of matrix operations are easier for triangular matrices. "Easier" here means that the time-complexity and memory requirements for a computer to calculate the result will be lower. If you need a lot of calculations on a matrix, obtaining its LU decomposition will likely speed things up.



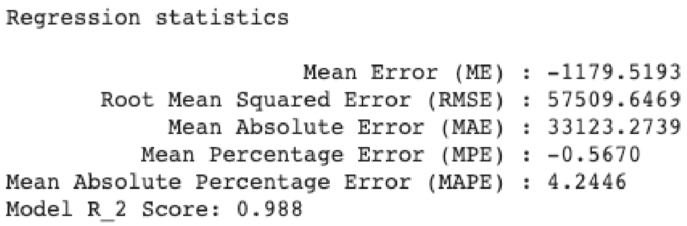
```
Regression statistics

                      Mean Error (ME) : -0.0000
       Root Mean Squared Error (RMSE) : 58430.9781
            Mean Absolute Error (MAE) : 33054.0824
          Mean Percentage Error (MPE) : -0.4248
Mean Absolute Percentage Error (MAPE) : 4.2476
Model R_2 Score: 0.987
```

Based on the regression statistics above, the model has a R^2-Score of 98.7%, implying that the model explains 98.7% of the variability in final price with respect to the feature set.

Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One or more variables are explanatory variables, and the remaining is the dependent variable.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX_i}$, where $\boldsymbol{X_i}$ is the "ith" explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).



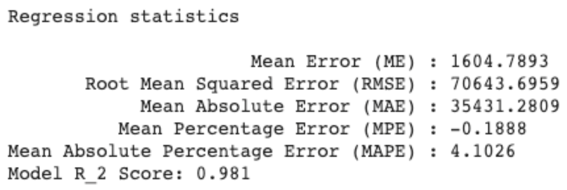
```
Regression statistics

                      Mean Error (ME) : -1179.5193
       Root Mean Squared Error (RMSE) : 57509.6469
            Mean Absolute Error (MAE) : 33123.2739
          Mean Percentage Error (MPE) : -0.5670
Mean Absolute Percentage Error (MAPE) : 4.2446
Model R_2 Score: 0.988
```

Based on the regression statistics above, the model has a R^2-Score of 98.8%, implying that the model explains 98.8% of the variability in final price with respect to the feature set.

Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class value (decision taken after computing all attributes). The paths from root to leaf represent regression rules.

Tree based learning algorithms are one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation.

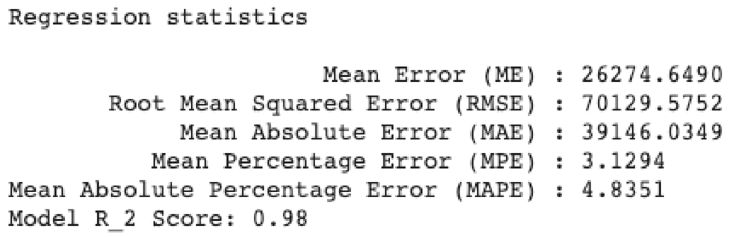


```
Regression statistics

                       Mean Error (ME) : 1604.7893
       Root Mean Squared Error (RMSE) : 70643.6959
            Mean Absolute Error (MAE) : 35431.2809
          Mean Percentage Error (MPE) : -0.1888
Mean Absolute Percentage Error (MAPE) : 4.1026
Model R_2 Score: 0.981
```

Based on the regression statistics above, the model has a R^2-Score of 98.1%, implying that the model explains 98.1% of the variability in final price with respect to the feature set.

Neural Network

Neural networks are the workhorses of deep learning. In this method, the model endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. While these networks are black boxes (their inner working is not explainable) they aim to accomplish the same thing as any other model — to make good predictions.

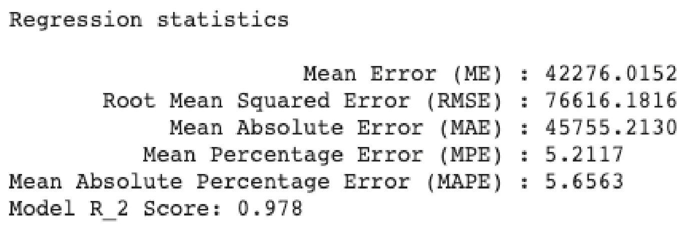


```
Regression statistics

                       Mean Error (ME) : 26274.6490
       Root Mean Squared Error (RMSE) : 70129.5752
            Mean Absolute Error (MAE) : 39146.0349
          Mean Percentage Error (MPE) : 3.1294
Mean Absolute Percentage Error (MAPE) : 4.8351
Model R_2 Score: 0.98
```

Based on the regression statistics above, the model has a R^2-Score of 98%, implying that the model explains 98% of the variability in final price with respect to the feature set.

Ensemble Method

Given that each model architecture is sensitive to different parameters, an ensemble method is applied to produce an aggregated output based on the four models used earlier. The simple average method is used in this case.

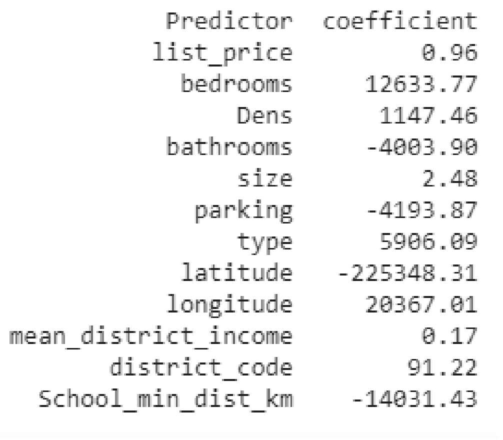


```
Regression statistics

                      Mean Error (ME) : 42276.0152
       Root Mean Squared Error (RMSE) : 76616.1816
            Mean Absolute Error (MAE) : 45755.2130
          Mean Percentage Error (MPE) : 5.2117
Mean Absolute Percentage Error (MAPE) : 5.6563
Model R_2 Score: 0.978
```

Using a simple average method, we obtain a model that has a R^2-Score of 97.8%, implying that the model explains 97.8% of the variability in final price with respect to the feature set.

Takeaways

At first glance, the model R^2 scores imply a near perfect prediction. However, it is important to analyze the individual features that comprise the dataset. Due to the black box nature of neural networks, it is not possible to perform meaningful feature analysis. Therefore, we conduct feature analysis on the first three models.

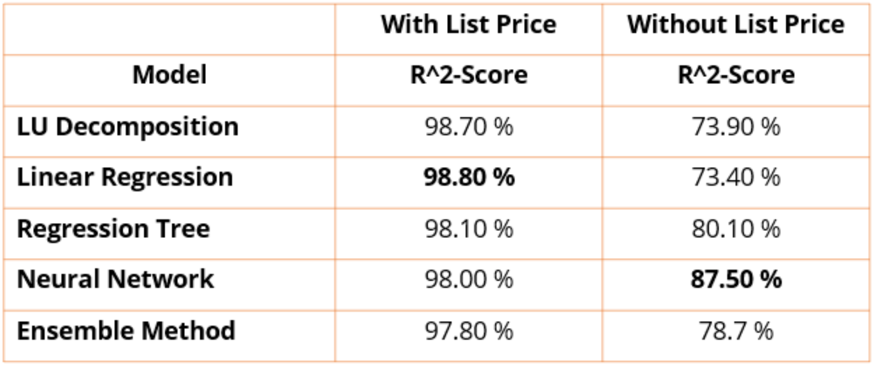


```
           Predictor  coefficient
          list_price         0.96
            bedrooms     12633.77
                Dens      1147.46
           bathrooms     -4003.90
                size         2.48
             parking     -4193.87
                type      5906.09
            latitude   -225348.31
           longitude     20367.01
mean_district_income         0.17
       district_code        91.22
     School_min_dist_km  -14031.43
```

Looking at the coefficients for list price and bedrooms, we see that the predictions are heavily influenced by the original asking price. Additionally, individual house features do not influence house prices significantly.

Taking a look at the coefficients for latitude and longitude we arrive at the conclusion that house prices increase as we move in the directions of North and West with respect to downtown Toronto, and decrease as we move in the directions of South and East.

It is important to remember that there is a segment of users that are not well versed with house pricing dynamics and therefore need to use our tool for price discovery. Therefore, we excluded the initial asking price as a feature, retrained our models and compared the results to our previously trained models. The following table summarizes the R^2 Score for the two sets of models:



	With List Price	Without List Price
Model	**R^2-Score**	**R^2-Score**
LU Decomposition	98.70 %	73.90 %
Linear Regression	**98.80 %**	73.40 %
Regression Tree	98.10 %	80.10 %
Neural Network	98.00 %	**87.50 %**
Ensemble Method	97.80 %	78.7 %

Comparing the two sets of models, we see an expected drop in R^2 Scores after removing the asking price feature from our training data.

Insights & Conclusion

Predicting house prices in Toronto for real estate transactions can be a complex process. We built four regression models using independent machine learning techniques, as well as an ensemble, that could input the cleaned data from two sources, and conduct a regression analysis to predict house prices with high accuracies. We faced a number of technical and business-related challenges when building these models but were able to finalize several insights on both counts. We were also able to visualize how these models could be easily deployed with an A.I. backend and a user-friendly interface.

Technical Challenges

There were three technical challenges we faced when building these models:

1. **Cleaning Data:** Making the data frames of our cleaned data so that it could be easily encoded across all four models, so that each model was dealing with the same set of features in a similar manner

2. **Model Evaluation:** We had some challenges with coming up with performance metrics for our ensemble methods
3. **Bias Detection:** We realized that our initial models were all very accurate and had to figure out the reason for this bias.

The technical insights we gained from these challenges was that it is easy to deploy multiple models for a regression problem, but it was very important to treat the data frames in a similar manner in each model. This was important if we wanted to compare the prediction performance across models. We also realized we could have made multiple models using each regression technique, by dropping /adding the features with equal weights from our exhaustive search, if we had more time. Interestingly, we also realized that the list / asking price feature we had used in all our models was making them very accurate, probably because there were human insights involved. Lastly, we came to the important understanding, that if all our models are doing relatively well, such as when our models had R^2 Scores close to 98% when using the list price feature, we don't really require an ensemble model. However, having the ensemble made more sense when we dropped the list price feature in our second iteration of the models.

Business Challenges

We had several challenges when trying to understand the business of trading houses. Our first big challenge was understanding what features or variables are important for predicting the price of a house (square-footage, location, schools nearby etc.). Our second challenge was finding reliable datasets on Toronto house prices as well as datasets with other features such as geographic data that we may need. Our third challenge was to decide whether we wanted to simply predict house prices or to also have other enhanced predictors. One such enhanced feature we discussed was whether we could also predict the duration a house would be on the market before it is sold given a certain list price.

We overcame our first two challenges by intensive research and taking expert help from our colleagues at Schulich's Real Estate Management program (MREI) as well as realtor friends. We decided not to go with the last idea of predicting time-on-market due to the time-series dependency of such a model. As the data we were provided was free of charge, we could only get a cross-sectional slice of what was on the MLS database on a certain date, and not data that would be suitable for a time-series analysis.

Furthermore, it also became clear to us that it was easier to look at certain models and explain the decision-making for clarity in a business-context. For example, the decision tree's decision-making process is easier to explain than the neural network model to an end-user. Lastly, we had to drop certain features, like house descriptions, which would have enriched our model's capacity to predict, due to our lack of NLP capabilities.

As for the major business insights we gained, firstly our models work better when we have the input on the initial asking price. In that, the list price was being set by someone with domain knowledge, so our models had an initial starting point which was close to the "minima", in this case, the "minima" being the price the house actually sold for.

Future A.I. Deployment & User-Interface

Having accurate machine learning-based predictive models for house prices is of not much use to the end-user who probably will be a non-data scientist, unless we automate the entire process and have a user-friendly and attractive interface. For our project to achieve this, we will first have to be able to automatically scrape the data from the data sources we used in real-time and feed it into a data frame that is readable by our model of choice. Next, we will need to have a user interface that has this option to get real-time data and be able to run the predictive machine smoothly, with just one-click.

Moreover, our user-interface will require to have filters that will enable them to predict house prices by location, property type, etc. As an advanced feature, we may also want to add the choice of using different variables to do the prediction. For example, the ability to add / drop the list price will allow naïve predictions which may be of use to owners of a new housing development or those flipping an abandoned / condemned home.

In summary, this project allowed us to use our knowledge of machine learning algorithms for a real-world business application. It taught us the value of finding relevant data sources, the need for diligent data pre-processing, and building and evaluating multiple regression models, including ensembles. Additionally, it gave us insights into the value of domain knowledge in enhancing predictive modelling. Lastly, although our project was on predicting home prices in Toronto, it drove home the relevance of using machine learning to increase the bottom-line of any business venture.