MSCI 446 – Project Proposal
Waterloo Rent Insight
Machine Learning Tool

Luke Doria - #20787989
Michael Lumibao - #20717846
Tonghe Bai - #20776084
Tony Cui - #20762110

1. Project Goals

The overall goal of the machine learning tool is to provide students with reliable insight regarding housing prices in the Waterloo region based on various factors.

After establishing an overall goal, different features or sub-goals of the housing prices tool can be outlined as follows. The first goal is the tool will be given a desired condition such as “bedrooms=2”, “freeParking=TRUE” and “busNearby=TRUE” and will output a price estimate for the unit. This should give the user a relatively accurate baseline of what the price should be for the unit after being trained on a large database of different listings. After the model outputs the relative price, it is simple for the user to compare with an actual price posting on an online listing to see if the posting is overpriced.

The secondary/optional feature is essentially the opposite of the first in which the user inputs an available budget for example “budget = 1000” and the model outputs approximate conditions that match the price range such as “bedrooms=1”, “freeParking=FALSE”. This gives the user a better understanding of whether his budget is sufficient for the different housing factors within the Waterloo region.

Through obtaining the project goals, a significant amount of business insight can be obtained such as which key features impact the price of housing in Waterloo. This would be very useful to example real estate companies when deciding on features that make a property more valuable to the public.

2. Problem Description and Importance

The process of renting housing is a well known problem for students at the University of Waterloo. There are a significant number of factors that impact students' decisions when searching for housing such as distance from the university, number of available rooms, physical square footage, nearby amenities, available parking and countless others making it extremely difficult to obtain a good combination. In addition, the COVID-19 pandemic has had a significant impact on pricing and availability of housing in the Waterloo region. According to a news article, in the spring of 2022 the overall rent price rose by more than 7% in the Waterloo region corresponding to the back to in person transition for the University [1]. Many landlords took advantage of the last minute transition to in person learning and raised the price of rent for desperate individuals searching for a place to rent for the school term. If individuals had a tool to determine the approximate price of rent based on previously stated features it would greatly assist in the house searching process. In addition, a tool would assist in combating the issue of landlords overcharging by giving tenants the insight into the true value of a rental property.

In addition, according to government research articles the overall rent in the Waterloo Kitchener area is up by 2.2% in the year 2020 and continues to increase. With rent prices increasing constantly, it is important to have a comparison tool for students to gauge different rent unit

prices to ensure that they are not overpaying. As a response to the growth in rent prices in the Waterloo area, it is expected that landlords would increase their rent per year. If a landlord increases their rent by a significantly unfair amount, the machine learning tool will be able to catch and prevent this to benefit the student population. Also, through research some average statistics and prices were found for different types of housing units within the Waterloo area as seen in the figure below.

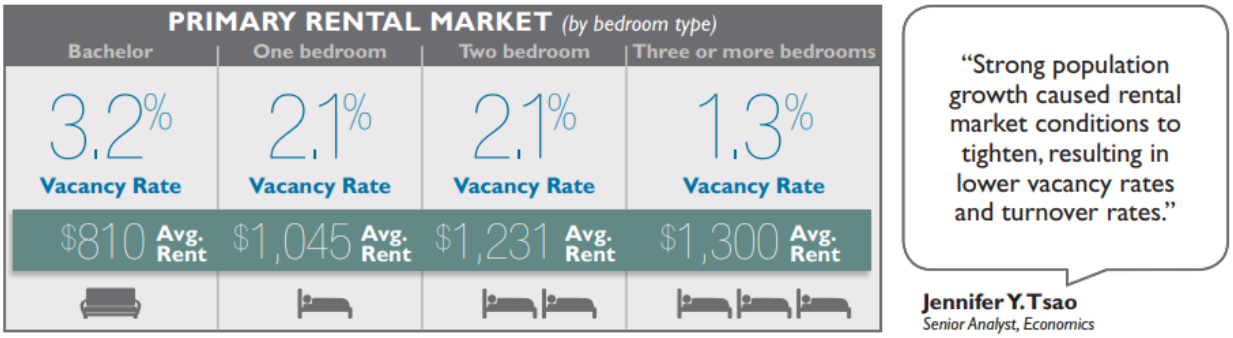


Figure 1: Vacancy Rate and Average Rent By Bedroom Type In Kitchener/Waterloo Region [2]

In conclusion, there is sufficient evidence that landlords have had a reason to unfairly raise rent for university students over the years in the Waterloo region and there is a significant need for a tool to counteract this from happening.

3. Dataset Description

In order for the machine learning algorithm to be successful in meeting the project goals and mitigating the problems of student housing, it is crucial that high quality data be obtained. Since all members of the group have been through the process of searching for University housing in Waterloo, it was decided that the best source would be from online postings on platforms such as the Facebook Marketplace or Kijiji based on experience. The only issue is that it would take an extremely long time to individually extract data points from single postings resulting in the group coming up with a clever automated solution. A publicly available web scraper from GitHub would be used to parse the online sources and extract the data from individual postings in a desired format with specifically chosen features. Another form of collecting data would be to use ChatGPT with a specific prompt to collect the data needed in JSON format. The data will eventually be collected in a spreadsheet or csv file, where each column is a feature and each row gives all the relevant information about a place for rent. Table 1 below shows the variable names of the most relevant features (the columns of the datasets), their data types, and their short descriptions.

Table 1: Summary of the Columns that Will Appear in the Dataset Obtained From the Webscraper

Category	Column	Data Type	Description
Location -related	dist_campus	float	The distance to the campus in km
	area	int	The area square footage
	groceries	bool	Whether there is easy access to grocery stores
	parking	bool	Whether parking is available
	transit	bool	Whether public transit is available
House-specific	num_rooms	int	The number of bedrooms in the unit
	num_bathrooms	int	The number of bathrooms in the unit
	hydro	bool	Whether hydro is included in the rent
	wifi	bool	Whether WiFi is available
	laundry	bool	Whether there are laundry and dryer machines within the unit
	furnished	bool	Whether the unit is fully furnished
	price	int	The monthly rent in Canadian dollars

The training dataset was decided to have a minimum size of 1000 rows, where each row gives all features of one place for rent. A sample dataset that contains three unique data points is provided below in the first 3 rows (Row 2 to Row 4) of the spreadsheet in Figure 2. The rest of the rows (Row 5 to Row 11) in Figure 2 are the data points collected by prompting ChatGPT and supplying it with the raw text of a listing page. The prompt supplied to ChatGPT can be found in the Appendix.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**dist_campus**	**area**	**groceries**	**parking**	**transit**	**num_rooms**	**num_bathrooms**	**hydro**	**wifi**	**laundry**	**furnished**	**price**
2	1.1	301	TRUE	TRUE	TRUE	1	1	FALSE	TRUE	TRUE	TRUE	1000
3	0.4	420	TRUE	FALSE	TRUE	2	1	FALSE	TRUE	FALSE	TRUE	1800
4	1.3	301	TRUE	FALSE	TRUE	1	1	FALSE	TRUE	FALSE	TRUE	850
5	N/A	N/A	TRUE	TRUE	TRUE	N/A	N/A	FALSE	TRUE	TRUE	FALSE	550
6	N/A	N/A	N/A	TRUE	N/A	7	3	TRUE	TRUE	N/A	TRUE	900
7	0.9	854	FALSE	TRUE	TRUE	3	2	TRUE	TRUE	FALSE	TRUE	986
8	0.4	N/A	FALSE	FALSE	TRUE	2	1	N/A	TRUE	TRUE	TRUE	1250
9	0.1	N/A	FALSE	N/A	TRUE	5	2	TRUE	TRUE	TRUE	FALSE	690
10	4.6	N/A	TRUE	TRUE	TRUE	N/A	1	FALSE	TRUE	TRUE	TRUE	1100
11	N/A	N/A	N/A	TRUE	TRUE	2	1	FALSE	TRUE	N/A	TRUE	808

Figure 2: Sample dataset that can be used for reference

By observing the sample dataset, the first thing noticeable is that the “area” field of many datasets is missing, which are labeled as “N/A”. For the 7 data points collected using ChatGPT, there is only one data point that gives the total area of the house. This implies that, when training the model using a larger dataset, it might be difficult to relate the “area” feature to the “price” attribute.

Apart from missing data, repeated data also tends to give less information about the correlations between a feature and the house price. For example, 9 out of 10 rows have “transit=TRUE” and all the 10 rows have “wifi=TRUE”, which renders the “transit” and “wifi” features less informative in terms of predicting the price. When building the project, one effective method that can be used to improve such cases is to re-define a particular attribute. For example, since public transit is available for the majority of these 10 places, a more meaningful way of using the public transit information is to specify the distance from the place to the nearest bus stop as numeric attributes (e.g. “50 m” rather than a “TRUE”).

For the numeric attributes, it can be observed in Figure 2 that “dist_campus” ranges from 0.1 km to 4.6 km, most places have less than 5 rooms and only 1 bathroom, and that the monthly rent varies from 550 to 1800 dollars, with most ranging between 770 and 1040 dollars. Similar patterns are expected to appear also in larger datasets that will be used when training the model. Histograms of the numeric data can be found in the Appendix.

The strategy will be to collect more data as necessary, so it is very likely that there will be more features used for the final implementation. After gathering data, if there is a feature that is missing from a lot of the rows, it may be dropped when training the model.

4. Machine Learning Algorithm

4.1 Problem Type

The proposed problem is a regression problem since the goal is to try to correlate various independent variables to a single variable, the cost of a housing rental unit. Defining such helps in determining what kind of model to use and what type of evaluation metrics to consider. The machine learning that will be implemented is numeric prediction, which predicts the rent of a residence based on its characteristics as outlined in the dataset description portion of the report.

In summary, when describing a machine learning problem, it is important to clearly define the type of problem, the features and class label, the dataset, the evaluation metrics, the algorithm or model, and the training and test datasets, which all are included in the following subsections.

4.2 Main Feature

The main feature of the proposed solution is to provide housing price prediction or justification. The algorithm will be given desired conditions as input (“num_rooms=2”, “parking=TRUE”, “transit=TRUE”, etc.), and output a price estimate. This should give the users, whether students or Waterloo housing agencies, a relatively accurate idea of what the price should be. Then the users may compare this pricing with the actual posted price to see if the posting is overpriced or reasonable to rent.

4.3 Class Label

This is the target or dependent variable that the team is trying to predict. In a regression problem, it is a continuous numeric value, specifically in this problem of predicting house prices, the class label would be the price represented in Canadian dollars as unsigned integers.

4.4 Dataset Justification

The data attributes presented earlier are sufficient to provide a decent estimation of the housing price based on the following reasons. The justification for each column is summarized in Table 2 below.

Table 2: Justification of dataset attributes

Category	Column	Justification
Location -related	dist_campus	The proximity of a house to a campus is an important factor that could influence the housing price.
	area	The size of the house is a crucial factor that determines its price. A larger house will generally cost more than a smaller

		one.
	groceries	The availability of grocery stores nearby is an important factor that could impact the housing price, as people prefer to live close to essential services.
	parking	Availability of parking spaces could also be a deciding factor for many renters and/or renters, and could influence the housing price.
	transit	Accessibility to public transportation is an important factor, as students prefer to live close to public transportation for convenience.
House-specific	num_rooms	The number of rooms in a house is a crucial factor in determining its price, as a house with more rooms is generally more expensive.
	num_bathrooms	The number of bathrooms in a house could also impact its price, as more bathrooms indicate more luxurious living.
	hydro	Whether hydro is included in the rent, like electricity, is an important factor that could impact the housing price, as people prefer houses with free access to basic amenities without the need in paying extras.
	wifi	Availability of free wifi facilities is an important factor in today's world, especially for students that rely on the internet for various academic or social activities.
	laundry	Availability of laundry facilities is an important factor that could influence the housing price.
	furnished	Whether the house is furnished or not is an important factor, as a furnished house is generally more expensive. Non furnished houses that require additional purchases are typically more affordable.
	price	Finally, the price of the house is what we are trying to predict as the class label, and it is dependent on all of the above factors.

4.5 Dataset Splitting

It is important to split the dataset into a training set and a test set, where the model is trained on the training set and evaluated on the test set. With a minimum of 1000 data entries or rows, the ratio that the group decides to use is 80:20, which designates 80% of the data as training and 20% as testing. In practice, other ratios like 70:30, 60:40, and even 50:50 are also employed.

4.6 Algorithm

It is clear that the examined housing price problem is a multivariable problem. The preliminary algorithm design is as follows.

The algorithm should start with 1-variable linear regressions, which would consider one independent variable at a time and fit a separate line for each independent variable. This should incorporate all combinations and investigate which of those may have a meaningful relation (i.e., number of bedrooms vs price, number of bedrooms vs number of washrooms). These 1 variable regression might be able to rule out some of the uncritical variables. We will have a high chance of finding that having free wifi or other variables may not really bring up the price and could be a disturbance to the multivariable fitting. Additionally, some research was conducted on the process of completing multi-variable linear regression for a dataset. The research indicated that for the subsequent 1-variable linear regressions, if the two independent variables have a large correlation coefficient of higher than 0.6 (positive and negative) then only one of them should be incorporated into the regression model [3]. Most of the analysis will be performed during the implementation phase of the project to rule out any correlated variables that were not foreseen in the project proposal.

Following such, the algorithm should perform multivariable piecewise linear regression. This would consider all of the independent variables simultaneously and fit multiple linear segments to the relationship between the independent variables and the dependent price variable. Piecewise linear regression should model the relationship as multiple linear segments, allowing for more flexible modeling of non-linear relationships.

The group is also considering doing 2 multivariable fittings. One being all variables vs price, the other being with only the significant variables vs price. Comparing the accuracy of results for both multivariable fitting should give the group insights on whether removing uncritical variables and simplifying the model will actually give a more accurate prediction.

4.7 Evaluation Metrics

It is common to evaluate the performance of a machine learning model based on various metrics. For example, in a regression problem, mean squared error or R-squared could be used, while in a classification problem, accuracy, precision, recall, or F1-score could be used. For a regression problem, there are no best metrics to use as it depends on the specific requirements of the problem and the characteristics of the data. The team is considering using multiple evaluation metrics to get a comprehensive understanding of the performance of a regression model.

Firstly, the Root Mean Squared Error (RMSE). It is the square root of the mean squared error and is expressed in the same units as the target variable. It is a more interpretable metric than the Mean Squared Error (MSE), as it provides a measure of the magnitude of the error in the

same units as the target variable and is also good for measuring the average magnitude of the errors.

Secondly, the R-squared. It measures the proportion of the variance in the dependent variable that is predictable from the independent variables, it indicates the goodness of fit of the model, or how well the model fits the observed data. Also by looking at the correlation coefficient R that varies between -1 and 1, the correlation of the regression model can be tested. The closer the value of R is to -1 or 1 represents data that is closely correlated and follows a set trend.

4.8 Optional Feature

Besides the main feature discussed above, the group is also considering implementing a housing recommendation feature based on available budget as input (“price=1000”). The output would be a list of housing attributes that approximately match with this price (“num_rooms=1”, “parking=FALSE”). Then the user may have a better understanding of whether the budget is sufficient for whatever condition they desire in the given market in Waterloo.

The challenge behind this is that the group may not be able to use the discussed regression method as we only have one single independent variable. The preliminary thoughts are to cluster the training data into various price ranges and assign tags to each cluster with the most frequent appearing housing attributes. An example would be for the cluster with pricing from $200-$300, a typical tag (representing the most likely condition) could be “no parking”. This optional feature may be withdrawn based on the progression of the main feature of the project.

5. Bibliography

[1] “Rents rose more than 7% in Waterloo Region between April and May,” *CambridgeToday.ca*. [Online]. Available: https://www.cambridgetoday.ca/local-news/rents-rose-more-than-7-in-waterloo-region-between-april-and-may-5482190. [Accessed: 08-Feb-2023].

[2] “Rental market report - publications.gc.ca.” [Online]. Available: https://publications.gc.ca/collections/collection_2020/schl-cmhc/NH12-83-2020-eng.pdf. [Accessed: 14-Feb-2023].

[3] R. Bevans, “Multiple linear regression: A quick guide (examples),” *Scribbr*, 15-Nov-2022. [Online]. Available: https://www.scribbr.com/statistics/multiple-linear-regression/. [Accessed: 14-Feb-2023].

Appendix

A: Parsing data using ChatGPT

Sample ChatGPT Prompt:

```
Can you parse these rooms for rent listings to extract the following data
in JSON format?

- dist_campus: distance to campus in kilometers.
- area: room size in square feet
- groceries: a boolean value if the location is near a grocery store like
walmart.
- parking: If parking is available
- transit: a boolean value If there is a bus stop or metro station nearby
- num_rooms: The number of rooms in the house
- num_bathrooms: The number of bathrooms in the house
- hydro: A boolean value if hydro is included
- wifi: A boolean value if wifi is included
- laundry: A boolean value if laundry is available
- furnished: A boolean value if the room is furnished
- price: The price of the room

Here is the post:

< Paste Listing here >
```

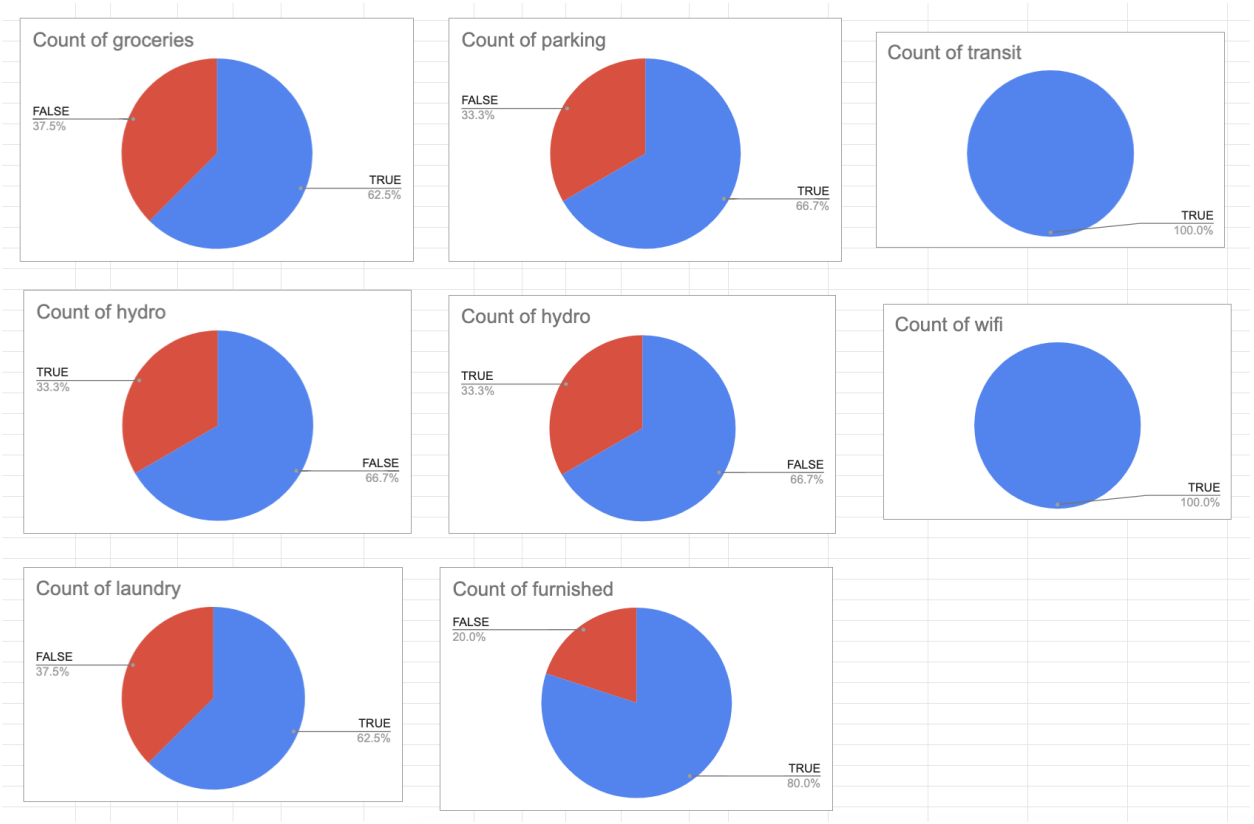
Sample ChatGPT's response:

```
Sure! Here's the JSON data extracted from the post:

{ "dist_campus": null, "area": null, "groceries": null, "parking": true,
"transit": true, "num_rooms": 2, "num_bathrooms": 1, "hydro": false, "wifi":
true, "laundry": null, "furnished": true, "price": 808 }

Note that some of the fields could not be parsed from the post, either because
the information was not provided or because it was not clear from the text.
```

B: Pie Chart of Boolean Features



C: Histogram of Numerical Features