University of Toronto

Faculty of Applied Science and Engineering

MIE 465 - Analytics in Action

Investigating the Factors that Affect the Nightly Price of Airbnb Listings

Final Report

Due: Monday, April 17 , 2017

Professor: Timothy Chan

Team 12

Nicholas Jalen Cheng	1001310761
Alex Kwan	1001559057
Ron Suprun	1000391444
Jiming Zhou	1001595393

1. INTRODUCTION

Airbnb hosts lack tools to determine an appropriate price for their space. However, using data from Airbnb listings, the Toronto Real Estate Board (TREB), as well as aggregated data about Toronto neighbourhoods, mathematical models have been constructed to discover which listing attributes have the most significant impact on a listing's nightly price. Identifying this relationship has lead to insights for potential hosts to predict their expected revenue.

2. DATA

The Airbnb listing data was sourced from the website "Inside Airbnb" [1] which scraped listings from the official Airbnb website. This data provided 95 attributes for each listing from Dec. 5th, 2016 and guest reviews. It used the San Francisco Model [2] to provide a conservative estimate of annual occupancy rates, capping the occupancy rate at 70%. As of Dec. 5, 2016 there are 12029 Toronto listings (average price $122) and 4728 Vancouver listings (average price $127).

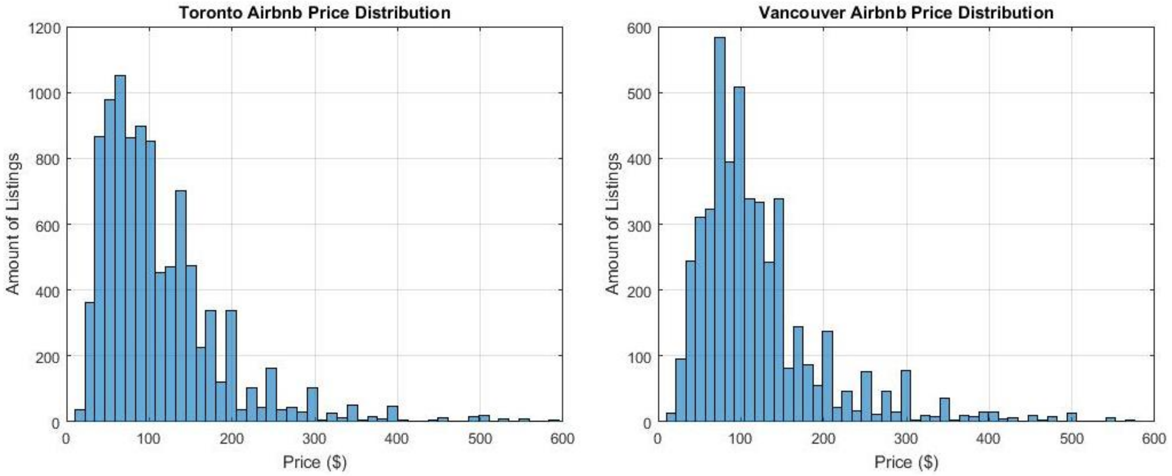


Figure 1: Comparison of Airbnb Listing Price Distribution in Toronto vs. Vancouver

Every quarter, the TREB listing Market Report [3] tracks lease transactions for condos in the GTA. The report provided the average monthly lease for 36 Toronto areas, broken down by number of bedrooms. The TREB area map [Appendix A] groups the 140 neighbourhoods into the 36 areas.

Toronto's diversity is captured in its demographic statistics for each neighbourhood. Tracked by the City of Toronto, these neighbourhood profiles [4] record the population, percentage of the population in each age group, average household income, and other useful measures.

3. METHODS

The two models constructed to predict nightly prices of Airbnb listings use linear regression and CART. They were built upon a combined dataset of Toronto Airbnb, TREB Rental, and Toronto neighbourhood data. The same methods were then applied on Vancouver Airbnb data (without the real estate and neighbourhood data) in order to test the robustness of the models. After data cleaning [Appendix B], 9937 Toronto listings and 3530 Vancouver listings remained. Then outlier listings with 0% availability, below 40% occupancy or a price over $1000 were removed.

The team started with linear regression because of its interpretability and flexibility to handle categorical and continuous variables. The predictor variable chosen was the listing price. Categorical feature variables were mapped to binary “dummy” variables [Appendix C].

The team used R^2 and root mean squared error (RMSE) to measure linear regression accuracy. R^2 indicates how much of the price variability is explained by the model, and provides an understanding of the model’s capability. RMSE measures the average difference between the predicted prices and actual prices. RMSE handles both positive and negative differences, is sensitive to small and large errors [5], and is easy to understand since it is in the same unit, of dollars, as the predicted variable, price [6].
The team also applied LASSO regularization to mitigate potential overfitting. The team chose 100 lambda values (iterations) since more than 100 lambdas did not improve the model accuracy [Appendix D]. Furthermore, 10 fold cross validation replaced 70/30 partitioning to provide an assessment of LASSO’s impact.

The linear regression model went through several iterations. Model 1 condensed the neighbourhoods into 6 general areas. For example, neighborhoods 12, 14, 16 etc belong to Downtown. Model 2 added LASSO regularization. Model 3 contained only the statistically significant attributes [Appendix E]. Model 4 contained only listings in neighborhoods with 100 or more listings. LASSO did not improve model accuracy, hinting the model does not suffer from overfitting, so it was dropped. Model 5 increased the location granularity to all 140 neighbourhoods in Toronto, allowing for the addition of neighbourhood rental data and demographic data such as population breakdown by age group and gender. Finally, Model 6 and 7 contain only statistically significant attributes with outliers removed [Appendix F].

The next tool used was a CART model. The CART model is favourable, because it captures nonlinear relationships and does not assume the dependent variable (price) is linearly related to the independent variables. Also the CART model is easy to interpret, and can make predictive splits on our highly categorical data. Another advantage of the CART model is that it automatically avoids splits on data deemed insignificant, while with linear regression, the manual removal of insignificant variables is necessary and difficult.

To reach the final most accurate CART model, several variations were built. Progressing through each of the following variations in Table 1, there was substantial accuracy improvement. Accuracy was judged using RMSE as well as 10%, 25%, and 50% tolerance levels with a minimum range of $10 plus or minus. For example if a listing actually costs $70, a prediction would be correct within a 10% tolerance if it predicted anywhere between $60 and $80.

Table 1: Comparison of Techniques used in CART Models

Model	Data	Validation/Partitioning Method
1	All listings with all attributes	70/30 training/testing sets
2	All listings with all attributes	10 trials of 10 fold cross validation
3	Outliers removed with all attributes	10 trials of 10 fold cross validation

Other CART models and Random Forest models that were developed but did not improve prediction accuracy are included in Appendix H.

As previously mentioned, moving from 70/30 training/testing sets to an industrial standard of the average from 10 trials of 10 fold cross validation, provided a better assessment of a model's accuracy. Cross validation also resolved issues with 70/30 partitioning, such as neighbourhoods not having enough listings in the training set after partitioning. Moreover, the average was taken since every time the CART model is run, different results are outputted due to random splits.

Originally bins were used [Appendix I] for the leaf nodes of the CART model, however there were consistency issues. For example if a listing's actual price is $99, and the bin it belonged to was $90-99, a $100 prediction would be deemed incorrect even though the prediction is only $1 off. Therefore tolerance ranges provide a consistent accuracy criteria. Furthermore, excluding outlier listings, or narrowing down listings based on an attribute resulted in better predictions.

4. ANALYSIS

Linear regression and CART results and discussion for both Toronto and Vancouver are given:

Table 2: Comparison of Linear Regression Model Accuracy - Toronto

#	Regression Model Progression	R^2	RMSE	Comment
1	Regular regression	0.264	132	

2	Add LASSO regularization	0.243	133.15	LASSO does not improve accuracy
3	Remove statistically insignificant feature variables	0.189	133.99	LASSO does not improve accuracy
4	Only listings in neighbourhoods with more than 100 listings	0.177	133.18	LASSO does not improve accuracy. Therefore the team did not overfit, so LASSO is removed
5	Regular regression with added neighbourhood rental and demographic data	0.285	131	
6	Remove statistically insignificant variables	0.258	133	
7	Outliers removed	0.504	66.3	Best model for Toronto

No overfitting is present as the model's accuracy is consistent across the different sized dataset versions and LASSO does not improve model accuracy. Model 4 and 6 show that adding demographic data did not improve model accuracy. Therefore neighbourhood demographics may affect long term tenants but does not influence short term tenants. Finally, removing outliers greatly improved model accuracy, since the nonsensical listings biases predictions.

Final Toronto linear regression model, which is Model 7 (See Appendix J for full model):
Price = -21.526 + 15.229(isChurchYongeCorridor) + 35.601(isLoft) … - 0.24793 (occupancy rate)

Table 3: Comparison of Linear Regression Model Accuracy - Vancouver

Model	Regression Model Progression	R^2	RMSE	Comment
A	Regular Regression	0.441	79.3	
B	Remove statistically insignificant feature variables	0.436	79.3	
C	Remove outliers	0.512	65.9	Best model for Vancouver
D	LASSO Regularization	0.458	67.5	Does not improve model accuracy, meaning the team did not overfit

Final Vancouver Linear Regression, which is Model D (See Appendix J for full model):

Price = -20.249 + 29.623(isDowntown) … - 4.5074(number of reviews per month)

One major finding is only a few variables have a statistically significant effect on the listing price. These variables include amenities such as Wi-Fi and a gym, only 1 neighbourhood in Toronto and only 6 neighbourhoods in Vancouver. This result is intuitive, since only certain neighborhoods are extremely lavish or convenient and only certain amenities are commonly desirable. Another finding is that the number of bathrooms and bedrooms, providing a flexible cancellation policy, and booking an entire property increases the listing price [Appendix J]. Furthermore, of all the property types, only loft significantly affects price. Lofts are seen as trendy and chic, provide a large space, and have high ceilings which are attractive for tourists [7]. Ultimately, linear regression predicts Airbnb listing prices with moderate accuracy. Figure 2 plots predicted price vs actual price for Toronto. For most listings with average prices, the model predicts it well. However linear regression struggles to predict prices that deviate far from the average price of $122.

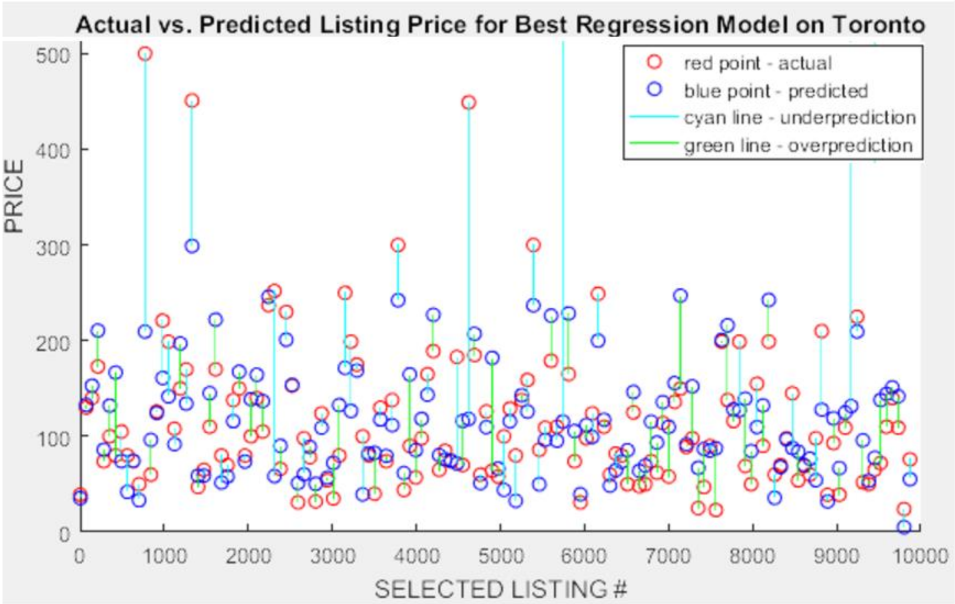


Figure 2: Actual vs Predicted Listing Price for Toronto (for Vancouver see Appendix K)

There are several reasons for why linear regression cannot achieve excellent accuracy. First, assumptions of normality and constant variance of residuals are not satisfied [Appendix L], which means linear regression is poorly suited for the data. Second, Airbnb hosts price their listing according to competition, demand, season, and location, so the price of one listing is not independent from the price others, violating the independence assumption. Third, non-linear relationships including the spike in price for certain neighbourhoods can not be captured.

Moving on to the CART, the results for model variations with significant improvement are given.

Table 4: Comparison of Toronto CART Model Accuracy by Tolerance Range (Appendix G)

Criteria:	10% Tolerance	25% Tolerance	50% Tolerance	RMSE
Model 1	130 splits, 33% correct	130 splits, 48% correct	30 splits, 81% correct	101.37
Model 2	370 splits, 33% correct	350 splits, 49% correct	410 splits, 81% correct	86.19
Model 3	350 splits, 35% correct	200 splits, 54% correct	400 splits, 85% correct	61.53

Model 3 is the final CART model that can be applied to all potential Toronto Airbnbs, and has a RMSE value of 61.53. Niches were explored to discover subsets of situations where the model is particularly accurate.

Table 5: Comparison of Toronto CART Model Accuracy at 25% Tolerance by Property Type

Criteria:	25% Tolerance Houses	25% Tolerance Townhouses	25% Tolerance Apartments	25% Tolerance Condos
Model 3	30 splits, 55% correct	270 splits, 43% correct	90 splits, 53% correct	40 splits, 51% correct

From Table 5, no particular property type is predicted better than the 54% average.

Table 6: Toronto CART Model Accuracy for Private Rooms by Tolerance Range

Criteria:	10% Tolerance Private Rooms	25% Tolerance Private Rooms	50% Tolerance Private Rooms
Model 3	30 splits, 51% correct	40 splits, 61% correct	30 splits, 91% correct

From Table 6, the 1,972 private room listings with minimum 40% occupancy, are predicted particularly well. Therefore, the 4000+ entire room/apartment listings with minimum 40% occupancy are predicted worse.

Looking at specific neighbourhoods: Annex, Dovercourt-Wallace Emerson Junction, Little Italy, South Riverdale, and Willowdale East stood out as neighbourhoods with over 100 listings, that were predicted better than the baseline 54% accuracy of Model 3, within 25% tolerance levels.

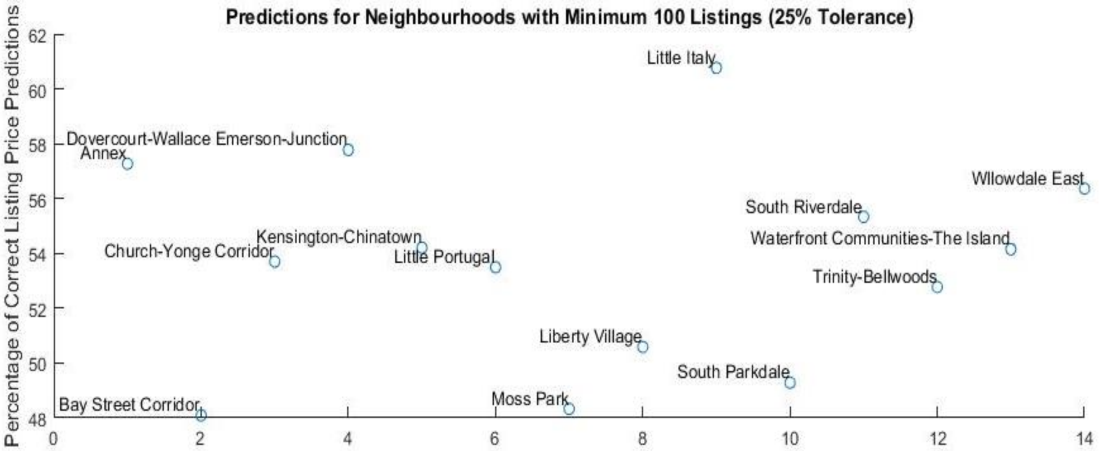


Figure 3: 25% Tolerance Accuracy for each Neighbourhood with >= 100 listings

Extending the CART Model 3 technique to Vancouver gave similar accuracy results.

Table 4: Comparison of Vancouver CART Model Accuracy by Tolerance Range (Appendix M)

Criteria:	**10% Tolerance**	**25% Tolerance**	**50% Tolerance**	**RMSE**
Model 3	120 Splits, 21% correct	80 Splits, 38 % correct	80 splits, 57% correct	57.97

The CART decision rules intuitively make sense. Splits at the top are most important and are made on the variables: room type, guests accommodated, and number of bathrooms which are significant variables in the linear regression model.

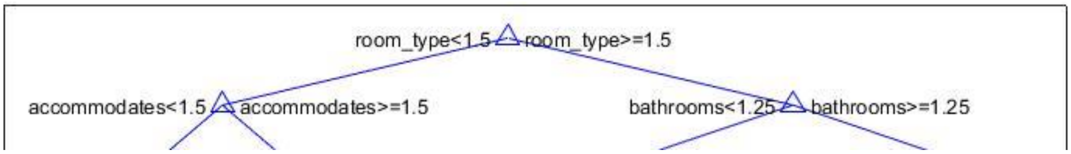


Figure 4: The top of CART Model 3

Overall, the CART model predicted price with moderate accuracy within 25% tolerance ranges, with approximately equal amounts of predictions being too high or too low.

Both the linear regression and CART model selected feature variables that suggest the size of the listing property is the most significant factor on price. The linear regression models assigned the highest coefficients to the variables isVilla, isEntireHome/Apt, number of bathrooms, and number of bedrooms, while the CART model had room type, number of guests, and number of bathrooms as the top splits. Comparing the performance of the two models using RMSE, the most accurate model is the CART model 3 with a RMSE of 61.53.

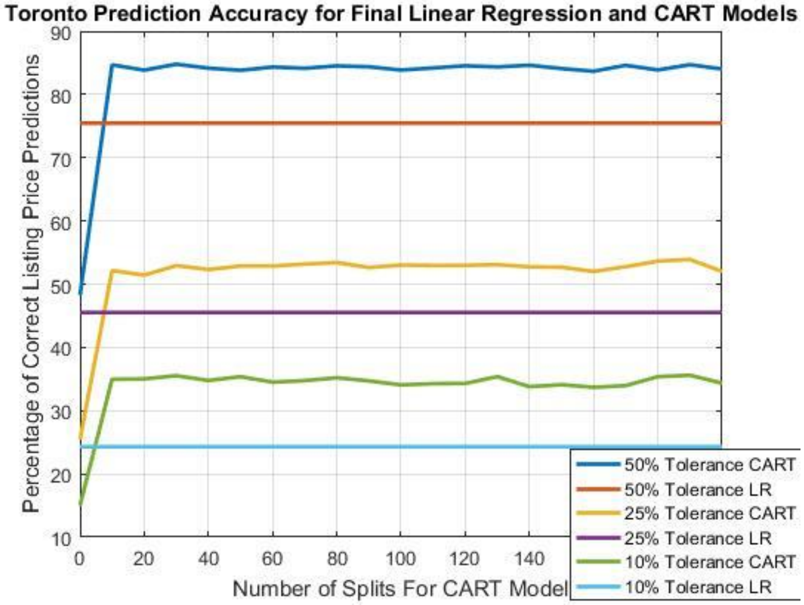


Figure 5: Tolerance Range Accuracy for the Final Linear Regression Model vs. Final CART Model

5. CONCLUSION

Overall, the most accurate model constructed to predict the nightly price of Airbnb listings in Toronto as well as Vancouver was the Model 3 CART model.

Attempts to combine results of linear regression and CART were not helpful, as choosing only the significant variables in the linear regression, as the only data for the CART model provided less accurate results than the CART model with all attributes.

Although the models' predictions were not very accurate, they do provide solid insights on the factors that affect nightly listing prices. Both models for both Vancouver and Toronto reveal that the most significant feature variables are related to either the location or the physical size of the listing. This observation means that listing price is closely linked to the value of the property being listed. This conclusion guides future model improvements, as more property value data should be incorporated into the models.

Also, since the size of an Airbnb listing is so significant in determining the listing price, separate models should be constructed for properties of each size. A factorial designed experiment should be performed to identify the impact of the categorical variables (property type, room type). Once a more accurate predictive model is developed, the focus of the project can return to the initial goal of helping hosts choose an optimal listing price. This model should consider the host's target market (families, single young adults, groups of young adults), historic tourism data in the host's city, and future seasonality trends such as sports, festivals and conventions.

REFERENCES

[1] "Inside Airbnb. Adding data to the debate.", Inside Airbnb, 2017. [Online]. Available: http://insideairbnb.com

[2] “About Inside Airbnb”, Inside Airbnb, 2017, [Online]. Available: http://insideairbnb.com/about.html

[3] ”Q4 listing Market Report”, Toronto Real Estate Board, 2016, [Online]. Available: http://www.trebhome.com/market_news/listing_reports/pdf/listing_report_Q4-2016.pdf

[4] ”Neighbourhood Profiles”, City of Toronto, 2017. [Online]. Available: http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=ae17962c8c3f0410VgnVCM10000071d60f89RCRD

[5] “MAE and RMSE - Which Metric is Better?”, Human in a Machine World, March 23 2016, [Online]. Available: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[6] “What are Mean Square Error and Root Mean Square Error?”, Vernier, 2011, [Online]. Available: https://www.vernier.com/til/1014/

[7] “The Pros And Cons Of Living In a Loft”, Stefan, 2014, [Online] Available: http://www.homedit.com/living-in-a-loft/

Appendix A: TREB Area Map

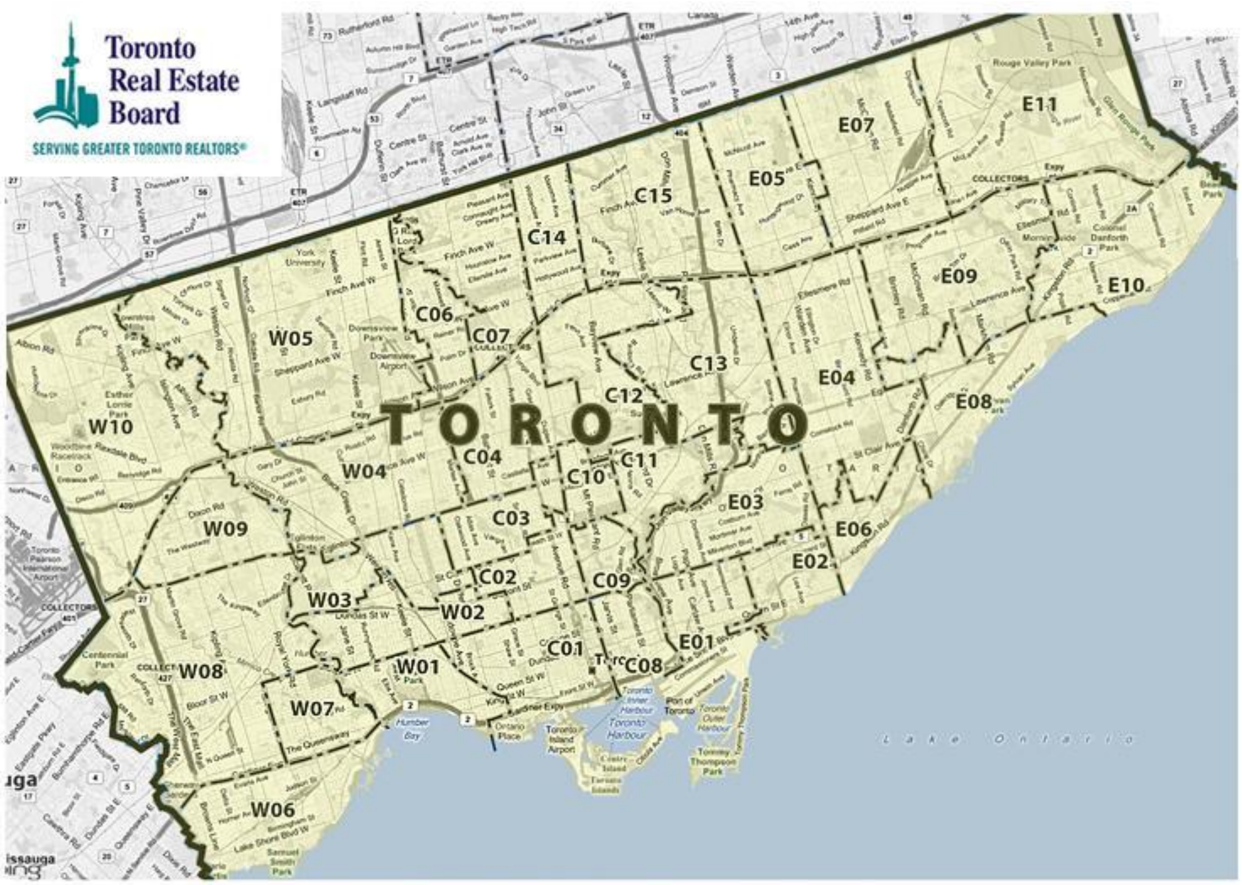
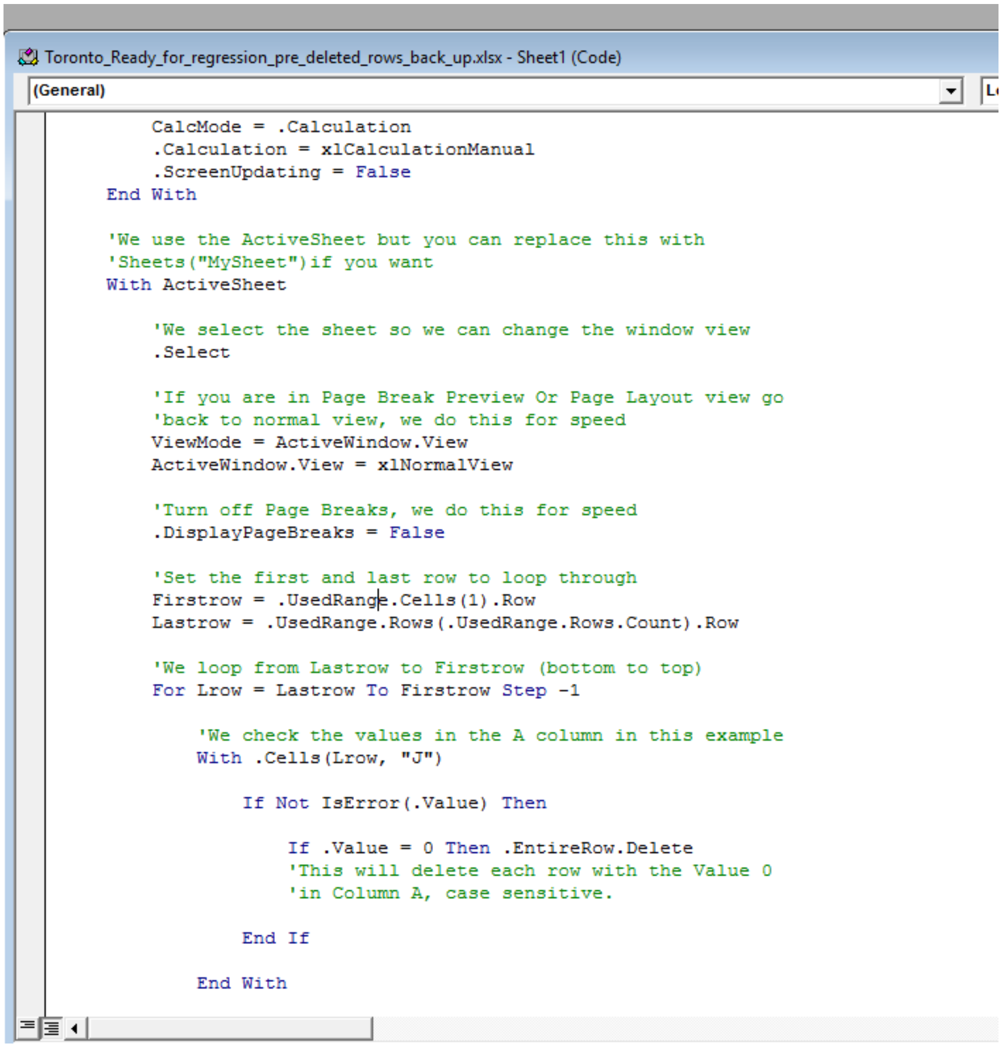


Figure 6: The 140 neighborhoods mentioned in the Inside Airbnb dataset were mapped to the 36 areas defined on the above map from the Toronto Real Estate Board.

Appendix B: Data Cleaning for Toronto and Vancouver

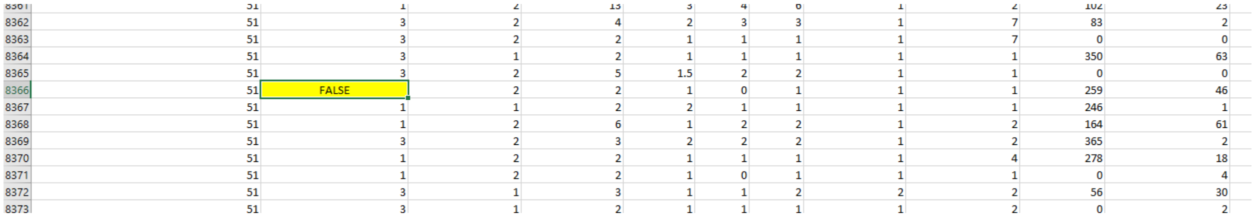
1. Deleted listings with 0 availability and 0 occupancy (i.e. listings that are not up for rental).


Toronto_Ready_for_regression_pre_deleted_rows_back_up.xlsx - Sheet1 (Code)
(General)
L

```
        CalcMode = .Calculation
        .Calculation = xlCalculationManual
        .ScreenUpdating = False
    End With

    'We use the ActiveSheet but you can replace this with
    'Sheets("MySheet")if you want
    With ActiveSheet

        'We select the sheet so we can change the window view
        .Select

        'If you are in Page Break Preview Or Page Layout view go
        'back to normal view, we do this for speed
        ViewMode = ActiveWindow.View
        ActiveWindow.View = xlNormalView

        'Turn off Page Breaks, we do this for speed
        .DisplayPageBreaks = False

        'Set the first and last row to loop through
        Firstrow = .UsedRange.Cells(1).Row
        Lastrow = .UsedRange.Rows(.UsedRange.Rows.Count).Row

        'We loop from Lastrow to Firstrow (bottom to top)
        For Lrow = Lastrow To Firstrow Step -1

            'We check the values in the A column in this example
            With .Cells(Lrow, "J")

                If Not IsError(.Value) Then

                    If .Value = 0 Then .EntireRow.Delete
                    'This will delete each row with the Value 0
                    'in Column A, case sensitive.

                End If

            End With
```

Figure 7: VBA code to delete all listing with zero availability and zero occupancy

2. Deleted uncommon property types, such as parking spaces and treehouses



8361	51	1	2	13	3	4	6	1	2	102	23
8362	51	3	2	4	2	3	3	1	7	83	2
8363	51	3	2	2	1	1	1	1	7	0	0
8364	51	3	1	2	1	1	1	1	1	350	63
8365	51	3	2	5	1.5	2	2	1	1	0	0
8366	51	FALSE	2	2	1	0	1	1	1	259	46
8367	51	1	1	2	2	1	1	1	1	246	1
8368	51	1	2	6	1	2	2	1	2	164	61
8369	51	3	2	3	2	2	2	1	2	365	2
8370	51	1	2	2	1	1	1	1	4	278	18
8371	51	1	2	2	1	0	1	1	1	0	4
8372	51	3	1	3	1	1	2	2	2	56	30
8373	51	3	1	2	1	1	1	1	2	0	2

Figure 8: Microsoft Excel was used to delete listings with uncommon property types, such as parking spaces and treehouses

3. Deleted attributes that are not important

The original data from Airbnb contained attributes such as id, host_id, latitude, and longitude etc. These were deemed not useful for data analysis, so they were removed.

Appendix C: Converting Categorical Variables Into Dummy Variables

For each categorical attribute (neighborhood, property type, cancellation policy etc), we represented it using one or more dummy variables.

For example, for cancellation policy = {strict, moderate, flexible}, we represented it using 2 dummy variables: isStrict, and isModerate. Note that flexible policy can be represented by setting both dummy variable to zero.

For Toronto's regression model #1, the team used 5 general areas instead of 140 neighbourhoods, this is how the dummy variables for it looked like:

- Is Downtown (is the listing located inside Toronto downtown area)
- Is Downtown East End
- Is Downtown West End
- Is Scarborough
- Is North York
- Is Etobicoke

Here is the dummy variables for property types:

- Is House (is the listing a house type)
- Is Townhouse
- Is Apartment
- Is Condominium
- Is Bungalow
- Is Bed & Breakfast
- Is Lofts
- Is Others

And the rest of dummy variables:

- IsPrivate = Dummy Variable, is the listing private (i.e not shared with a different tenant)?
- IsEntire = Dummy Variable, are you renting out the entire unit (e.g renting out the entire house)?
- IsRealBed = Does the listing provide a real bed (i.e some listing provide a couch as a bed)?

- IsStrict = Does this listing have a strict cancellation policy?
- IsModerate = Does this listing have a moderate cancellation policy?
- A1 = Dummy variable, does the listing have a TV
- A2 = Dummy variable, does the listing have a cable TV
- A3 = Dummy variable, does the listing have internet
- A4 = Dummy variable, does the listing have wireless internet
- A5 = Dummy variable, does the listing have air conditioning
- A6 = Dummy variable, does the listing have a pool
- A7 = Dummy variable, does the listing have a kitchen
- A8 = Dummy variable, does the listing have free parking on premises
- A9 = Dummy variable, does the listing allow smoking
- A10 = Dummy variable, does the listing allow pets
- A11 = Dummy variable, does the listing have a doorman
- A12 = Dummy variable, does the listing have a gym
- A13 = Dummy variable, does the listing have a breakfast
- A14 = Dummy variable, does the listing have pets living in it
- A15 = Dummy variable, does the listing have a elevator
- A16 = Dummy variable, does the listing have a hot tub
- A17 = Dummy variable, does the listing have a indoor fireplace
- A18 = Dummy variable, does the listing have a buzzer/wireless intercom
- A19 = Dummy variable, does the listing have heating
- A20 = Dummy variable, is the listing family/kid friendly
- A21 = Dummy variable, is the listing suitable for events
- A22 = Dummy variable, does the listing have a washer
- A23 = Dummy variable, does the listing have a dryer
- A24 = Dummy variable, does the listing have a smoke detector
- A25 = Dummy variable, does the listing have a CO detector
- A26 = Dummy variable, does the listing have a first aid kit
- A27 = Dummy variable, does the listing have a safety card
- A28 = Dummy variable, does the listing have a fire extinguisher
- A29 = Dummy variable, does the listing have essentials amenities
- A30 = Dummy variable, does the listing have a shampoo
- A31 = Dummy variable, does the listing have 24 hour check in
- A32 = Dummy variable, does the listing have cloth hangers
- A33 = Dummy variable, does the listing have a hair dryers
- A34 = Dummy variable, does the listing have an iron
- A35 = Dummy variable, does the listing have a laptop friendly workspace
- A36 = Dummy variable, does the listing have a lock on bedroom door

The team applied similar methodology for Vancouver.

Appendix D: Determining Number of Iterations (Lambdas) in Linear Regression with LASSO

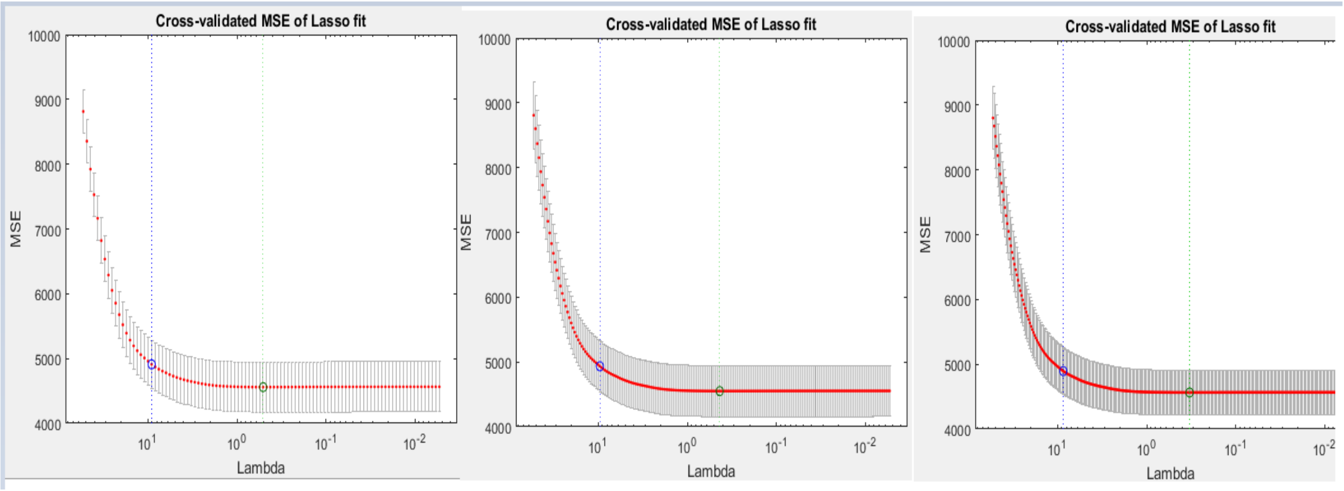


Figure 9: 10 fold cross validation with 100, 200, and 300 iterations all produce similar results (in this case, around MSE = 4800). Hence doing 100 iterations is good enough

Appendix E: Identifying Statistically Significant Attributes

From Toronto’s linear regression model #2 to #3, and from Vancouver’s linear regression model #1 to #2, the team reran the linear regression on only statistically significant variables, defined by **having a pValue of less than 0.05**. This was done in order to make the model easier to interpret, and hopefully more accurate. Below is an example of the process, with Vancouver data.

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-11.679	16.531	-0.70652	0.4799
neighbourhood_2	32.573	14.763	2.2064	0.027404
neighbourhood_3	17.821	15.229	1.1702	0.24198
neighbourhood_4	14.591	16.102	0.90617	0.36489
neighbourhood_5	9.4664	15.101	0.62686	0.53078
.				
.				
.				
property_type_3	-3.9782	3.6406	-1.0927	0.27457
property_type_6	1.9934	12.539	0.15898	0.87369
property_type_7	56.912	12.738	4.468	8.0817e-06
property_type_8	27.14	9.2338	2.9392	0.0033069
.				
.				
.				
availability_365	0.032926	0.0093932	3.5053	0.00046038
number_of_reviews	0.05869	0.062477	0.93938	0.34758
.				
.				
.				

Significant feature variables are highlighted in red. Eliminating the insignificant variables, the new linear regression model looks like this:

$\mathbf{Y} = \boldsymbol{\beta}_1(\text{neighbourhood\_2}) + \boldsymbol{\beta}_2(\text{property\_type\_7}) + \boldsymbol{\beta}_3(\text{property\_type\_8}) + \boldsymbol{\beta}_4(\text{availability\_365}) + \ldots\ldots$

Appendix F - Removing Outliers for Linear Regression

The team removed outliers using Cook's distance. Here is the MATLAB code:

```
%cook's distance
plotDiagnostics(mdl,'cookd') %Cook's distance

%Identify the outlier
[~,larg] = max(mdl.Diagnostics.CooksDistance)
T(larg,:)
```

Also for both linear regression and CART, listings with availability of 0% or an occupancy rate below 40% were removed, leaving 6,053 listings.

Appendix G: Significant CART Models

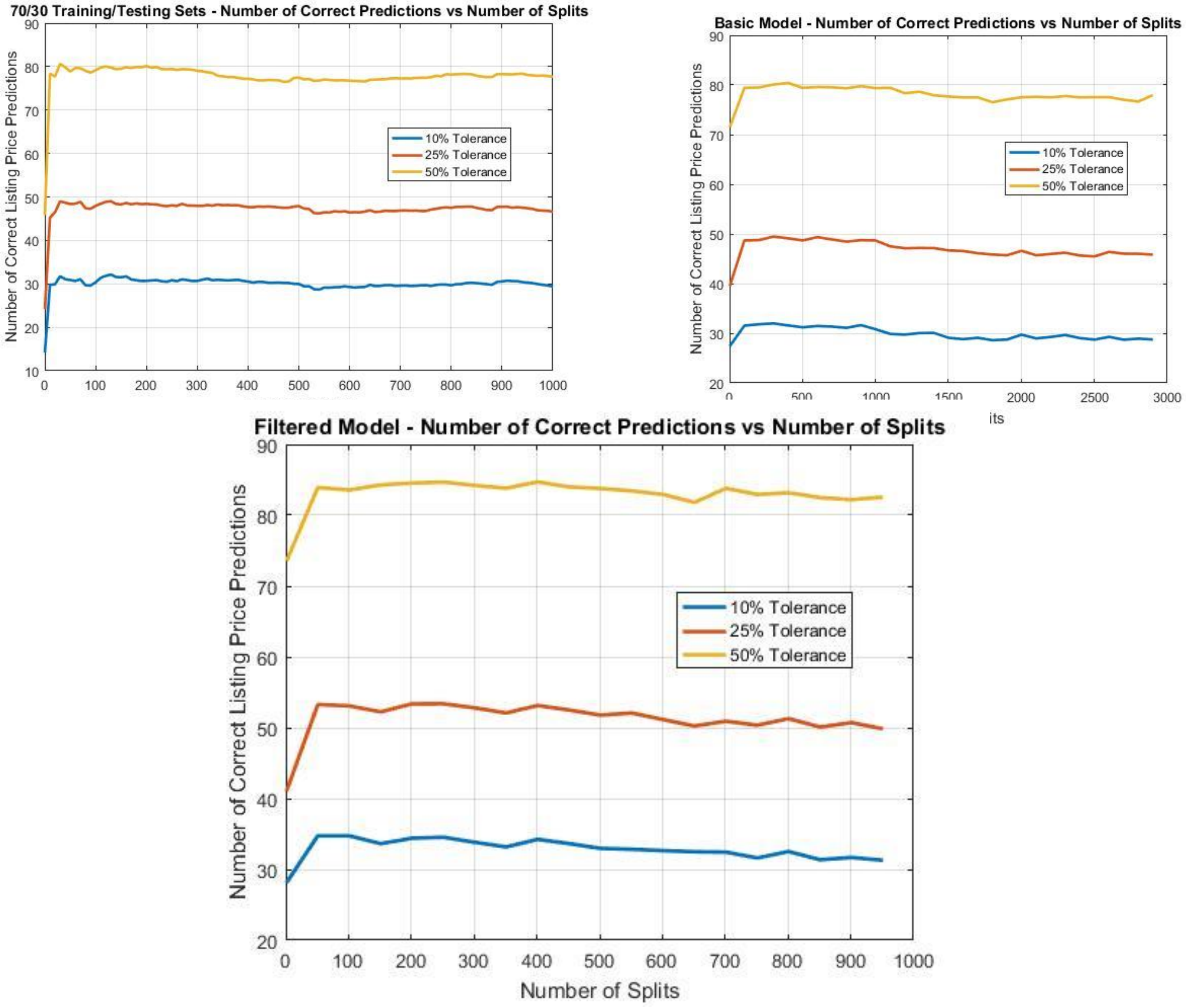
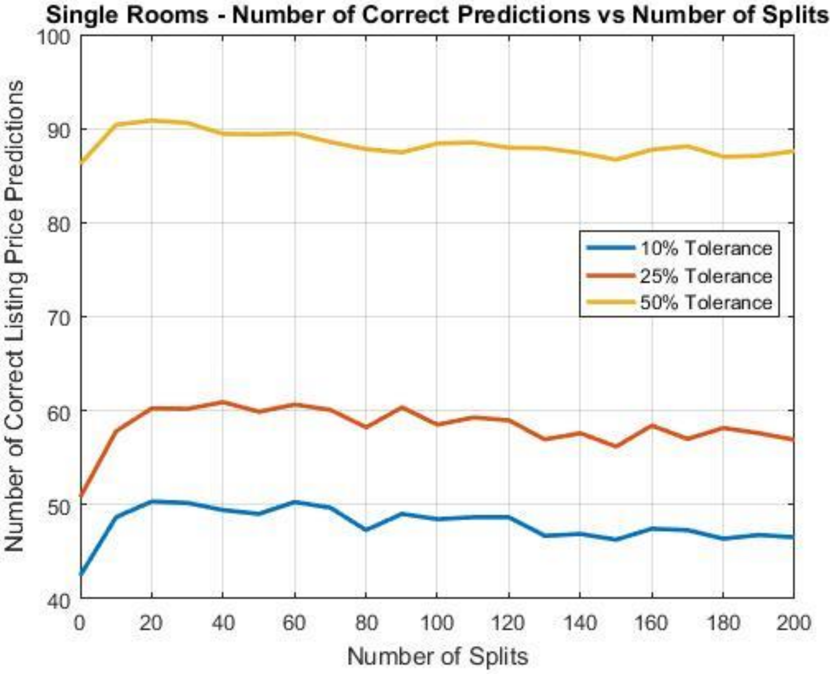
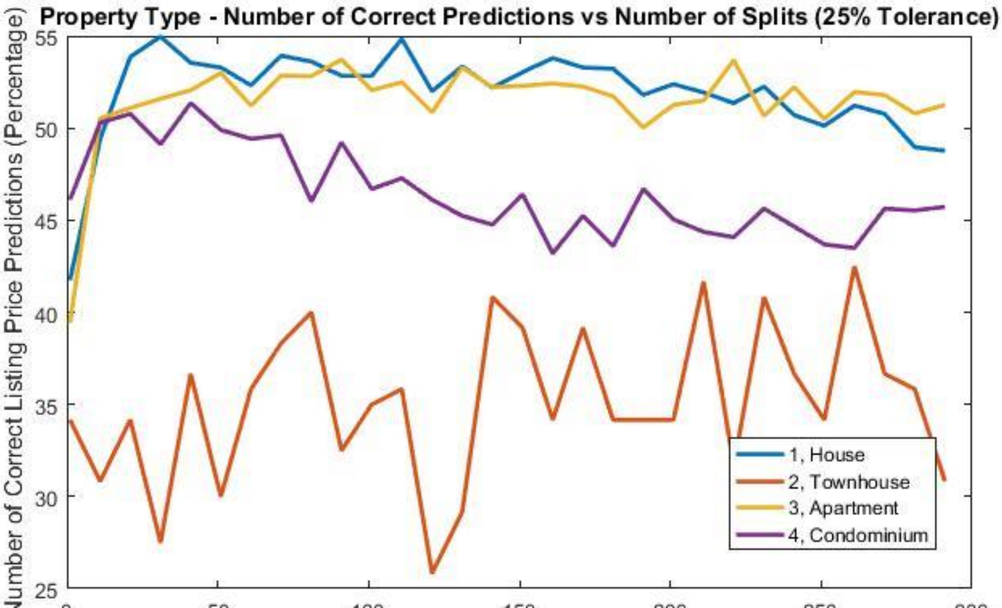


Figure 10 - Additional CART models that did not improve prediction accuracy

Appendix H: Other CART Models



Appendix I: CART Model Tolerance Definition



Grouping Number ($10 bins)	0% Tolerance (±Groupings)	Ranges ($)	25% Tolerance (±Groupings)	Ranges ($)	50% Tolerance (±Groupings)	Ranges ($)
1	±0	0 to 9	±0	0 to 9	±0	0 to 9
2	±0	10 to 19	±0	10 to 19	±1	0 to 29
3	±0	20 to 29	±0	20 to 29	±1	10 to 39
4	±0	30 to 39	±1	20 to 49	±2	10 to 59
5	±0	40 to 49	±1	30 to 59	±2	20 to 69
6	±0	50 to 59	±1	40 to 69	±3	20 to 89
7	±0	60 to 69	±1	50 to 79	±3	30 to 99
8	±0	70 to 79	±2	50 to 99	±4	30 to 119
9	±0	80 to 89	±2	60 to 109	±4	40 to 129
10	±0	90 to 99	±2	70 to 119	±5	40 to 149
11	±0	100 to 109	±2	80 to 129	±5	50 to 159
12	±0	110 to 119	±3	80 to 149	±6	50 to 179
13	±0	120 to 129	±3	90 to 159	±6	60 to 189

Appendix J: Final Linear Regression Models

Toronto:

Price = -21.526 + 15.229(isChurchYongeCorridor) + 35.601(isLoft) + 111.04(isVilla) + 43.342(isEntireHome/Apt) - 14.533 (isSharedRoom) + 7.8653(number of accommodates) + 31.225(number of bathrooms) + 28.109(number of bedrooms)+ 0.044367(availability per year) + 0.41171 (cleaning fee) + 9.2206(ifCancellationPolicyFlexible) +26.048(hasGym) + 9.5687(hasBreakfast) + 14.606(hasIndoorFireplace) - 4.1792(is family friendly) + 7.8512(listing suitable for events) - 0.24793 (occupancy rate)

Vancouver:

Price = -20.249 + 29.623(isDowntown) - 23.891(isKensingtonCedarCottage) -29.474(isMarpole) -28.664(isRenfrewCollingwood) - 33.196(isSunset) -36 (isVictoriaFraserview) + 30.268(isLoft) + 23.396(isTownhouse) + 36.637(isEntireHome/Apt) + 13.038 (isSharedRoom) + 7.5321(number of accommodates) + 22.788(number of bathrooms) + 36.942(number of bedrooms) + 0.039291(availability per year) + 0.017193(security deposit amount) + 0.27468 (cleaning fee) + 7.1904(ifCancellationPolicyFlexible) -11.963(require guest phone verification) - 4.5074(number of review per month)

Appendix K: Actual vs Predicted Listing Prices for Best Linear Regression Model on Vancouver

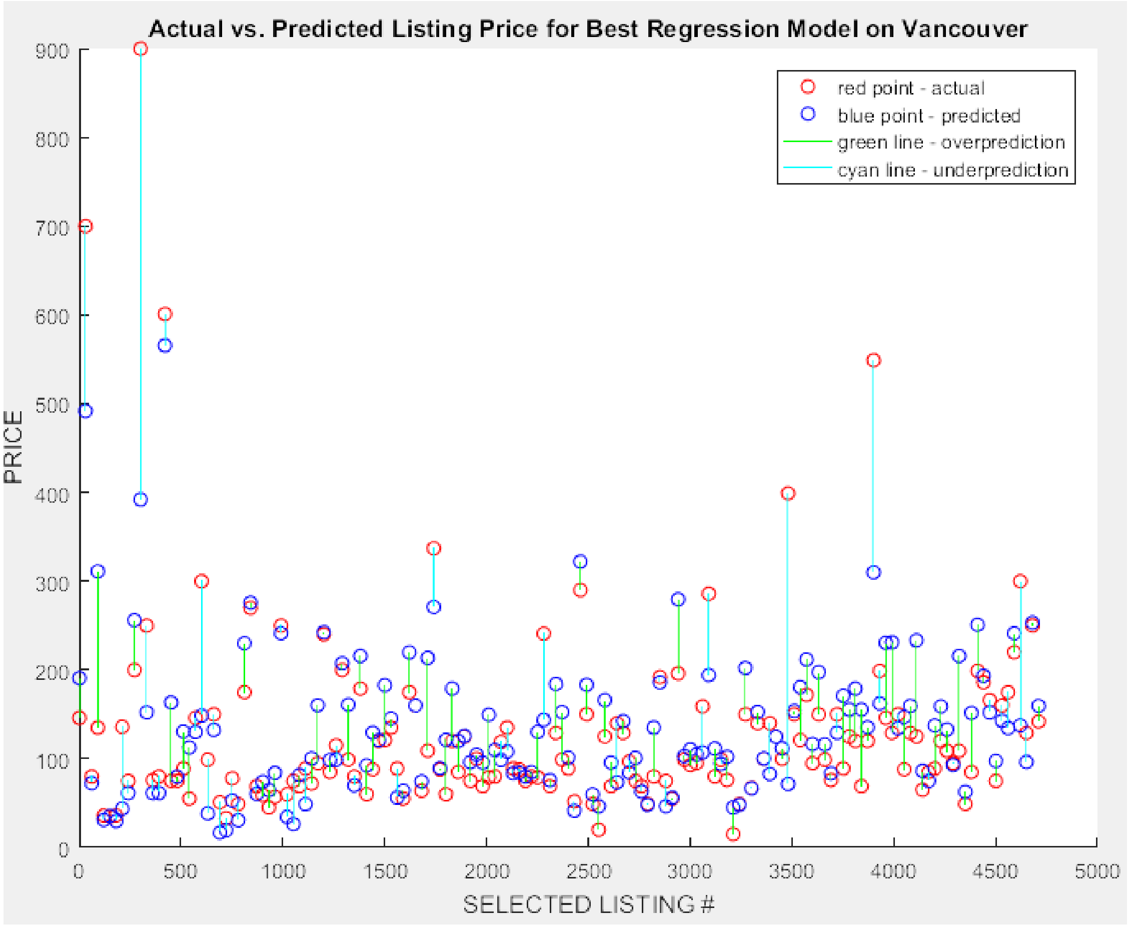
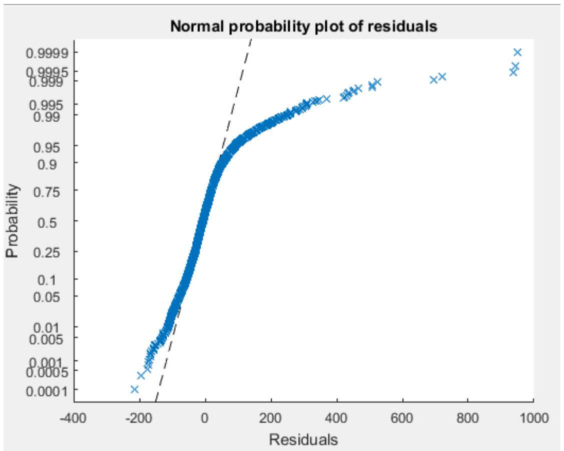
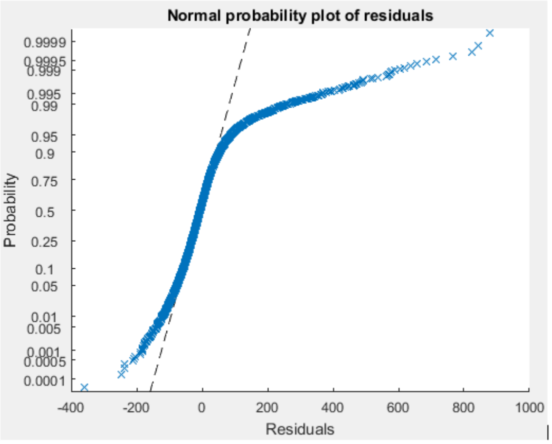
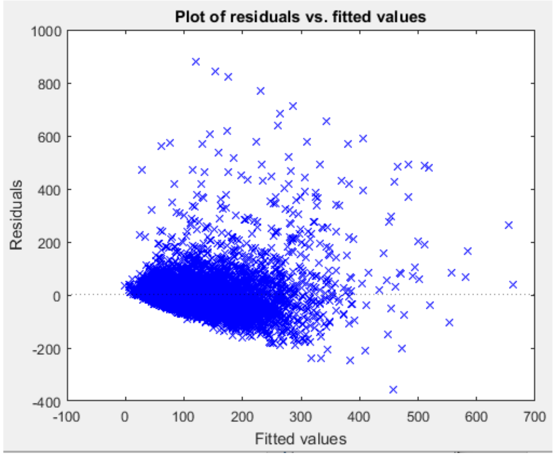
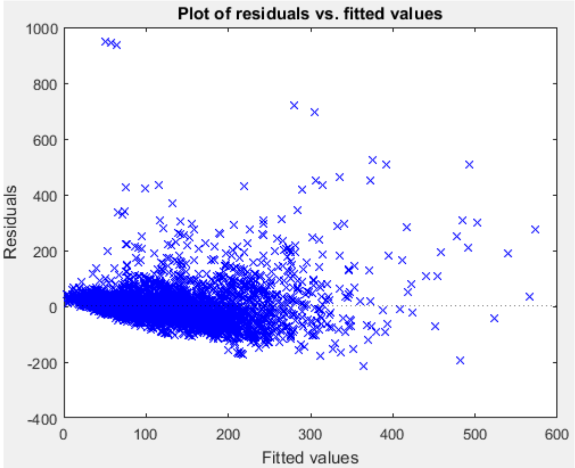


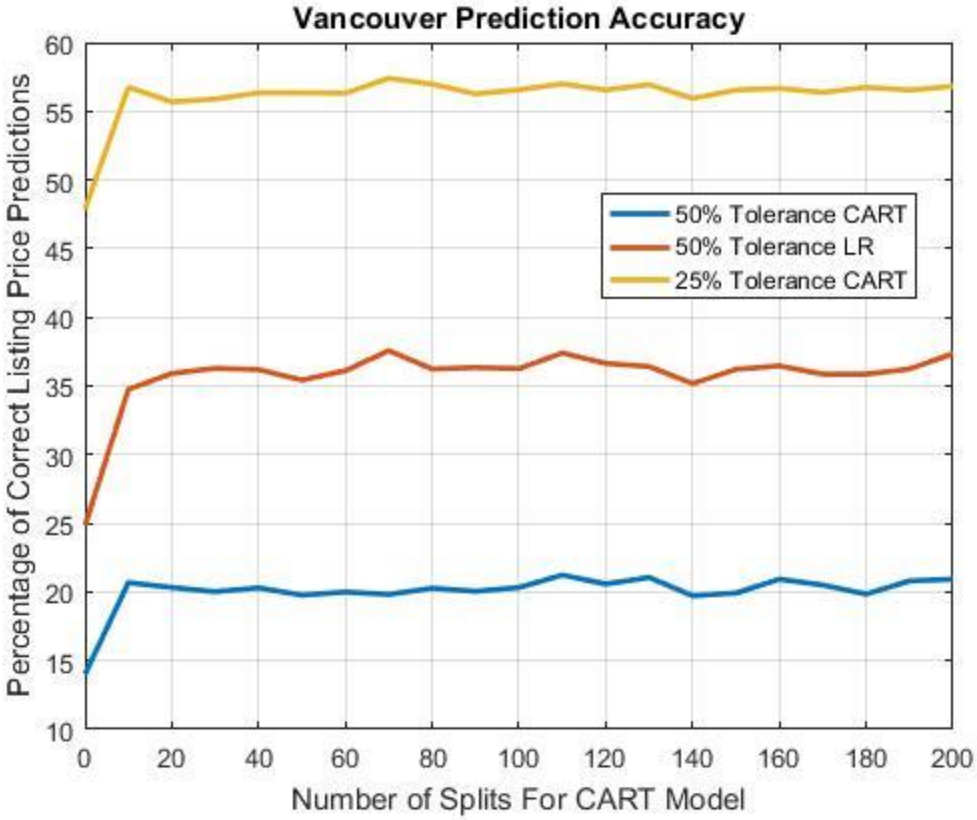
Figure 10 - The distances between predicted and actual prices are closer for listings whose prices are near the average, which is similar for Toronto.

Appendix L: Linear Regression Model Assumptions

The following plots reflects the fact that the Toronto and Vancouver data sets do not satisfy the assumption of normality and the assumption of equal residual variance.

	Toronto	Vancouver
Normality		
Equal Variance		

Appendix M: Vancouver CART Model Results


Vancouver Prediction Accuracy
Percentage of Correct Listing Price Predictions
Number of Splits For CART Model
50% Tolerance CART
50% Tolerance LR
25% Tolerance CART