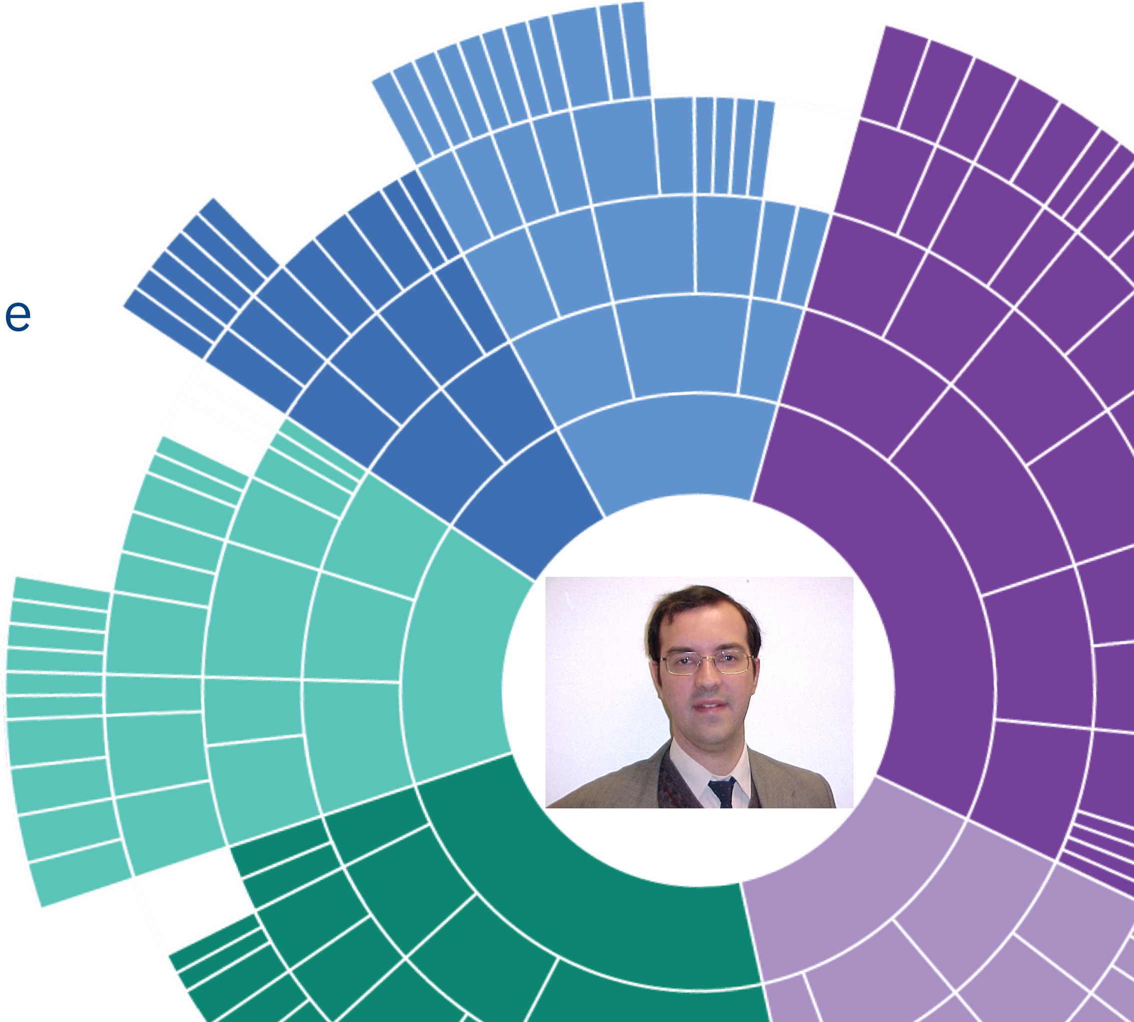
Applied Data Science Capstone Final Assignment Report
Bernard F.

October 2018



Introduction

Business Drivers:

- Canada: significant increases in real estate prices in large cities.
- Social issues
  - Migrations of younger people

Project Goals:

- Leverage the exploration of Toronto neighborhoods
- Add data about house prices
- Find relationship between prices and the types of venues near the properties being sold
- Predict property value based on the types of venues available (or not) nearby.

Target Audience:

- Real estate agents in the Toronto area
- Anyone planning on buying a house or investing in real estate in the area

Data

Ontario Property Sales

- Public domain data
- Features:
  - (unnamed): index
  - Address: Street address of the property in question
  - AreaName: Neighborhood where the property is located
  - Price ($): Selling price of the property
  - lat: Latitude
  - lng: Longitude
- Sample:

	Address	AreaName	Price ($)	lat	lng
0	86 Waterford Dr Toronto, ON	Richview	999888	43.679882	-79.544266
1	#80 - 100 BEDDOE DR Hamilton, ON	Chedoke Park B	399900	43.25	-79.904396
2	213 Bowman Street Hamilton, ON	Ainslie Wood East	479000	43.25169	-79.919357
3	102 NEIL Avenue Hamilton, ON	Greenford	285900	43.227161	-79.767403
6	#1409 - 230 King St Toronto, ON	Downtown	362000	43.651478	-79.368118
7	254A Monarch Park Ave Toronto, ON	Old East York	1488000	43.686375	-79.328918
8	532 Caledonia Rd Toronto, ON	Fairbank	25	43.691193	-79.461662

Data

FourSquare Venue Data

- Obtained through public API, free account
- Key Features:
  - Latitude, Longitude, Distance, Category
- Sample:


```
{'meta': {'code': 200, 'requestId': '5bce02af9fb6b75291665634'},
 'response': {'groups': [{'items': [{'reasons': {'count': 0,
       'items': [{'reasonName': 'globalInteractionReason',
         'summary': 'This spot is popular',
         'type': 'general'}]},
      'referralId': 'e-0-4ff1dbf1e4b07cca845d6e91-0',
      'venue': {'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/ju
icebar_',
           'suffix': '.png'},
          'id': '52f2ab2ebcbc57f1066b8b41',
          'name': 'Smoothie Shop',
          'pluralName': 'Smoothie Shops',
          'primary': True,
          'shortName': 'Smoothie Shop'}],
       'id': '4ff1dbf1e4b07cca845d6e91',
       'location': {'address': '265 Wincott Drive, Unit 2A',
        'cc': 'CA',
        'city': 'Etobicoke',
        'country': 'Canada',
        'distance': 56,
        'formattedAddress': ['265 Wincott Drive, Unit 2A',
         'Etobicoke ON M9R 2R7',
         'Canada'],
        'labeledLatLngs': [{'label': 'display',
          'lat': 43.67952707,
          'lng': -79.54477308}],
        'lat': 43.67952707,
        'lng': -79.54477308,
        'postalCode': 'M9R 2R7',
        'state': 'ON'},
       'name': 'Booster Juice',
       'photos': {'count': 0, 'groups': []}}},
```

Methodology

Data Understanding

Ontario Property Sales:

- Examined using Excel
- Confirmed the data types were appropriate (.dtypes)
- Descriptive summary using describe()
- Data quality issues:
  - Missing the area names
  - Zero as the sale price
    - Were those properties given away for free?
  - Latitude and/or longitude set to -999
  - Data set is bigger than just Toronto
  - Duplicate entries

FourSquare Venue Data:

- No issues were found.

Methodology

Data Preparation

Addressing Data Quality Issues

- Missing area names: do nothing
- Zero sale price:
  - Determine the scope: histogram, distribution plot of bottom 20% of the data, line plots and violin plots to confirm.
  - Significant number of records with suspiciously low values.
  - Addressed by excluding the bottom 1000 sale prices from the data set.
  - Suspiciously high values (top 1000) excluded as well.
- Invalid latitudes and longitudes (-999):
  - Geolocation data is key.
  - Small number of observations affected (153 out of 20,000+).
  - Visually inspected.
  - Dropped from the data set.
- Properties outside Toronto:
  - Found Toronto coordinate.
  - Added a feature representing the distance of any given property to that coordinate.
  - Used radius of 20 Km as a cut-off.
  - Eliminated from consideration all properties further away than that.
  - Just over 4,000 observations remained.
- Duplicate observations: dropped these observations.

Methodology

Data Preparation

Additional Steps

- Use FourSquare API to find venues close to the properties:
  - Close = "within 200 metres” (comfortable walking distance)
  - Constraint: 950 regular API calls / day
  - Sampling (size = 200)
  - Visually verified the distribution of properties on the map of Toronto
- List of venues:
  - Over 6,000 venues
  - Over 300 categories
  - A couple of properties did not have any nearby venues (dropped)
- Change the text features to numeric values
  - One-hot encoded using get_dummies()
  - Grouped the observations for each address using the mean value of each feature
  - Merged the resulting dataframe with the original

Methodology

Modelling and Evaluation

- Goal = predict the property selling price
  - Regression problem
- Combination of features as predictors
  - Multivariate regression
- Linear Regression model
  - Split sample data in training and test sets
  - Fit the Linear Regression model
  - Evaluation with Cross Validation (folds = 10)
  - Poor results
- Permutation Importance:
  - Intent: facilitate future studies
  - Identified 15 features that most affected the model
  - Refit and cross-validation: accuracy improved, but only marginally.
- Shapley Additive Explanation (SHAP) values:
  - Understand how those 15 features were affecting the model
  - Positive or negative effect of each feature
  - SHAP summary plot

Results and Discussion

Results:

- Unable to find a method to predict property prices in Toronto solely based on the nearby venues
- Larger number of observations not likely to improve results

Discussion:

- Data set used was not complete enough
- Additional features, such as property size, should result in a much more accurate model.
- Using the Area Name feature
- Recommendation: future studies
  - May use the features I identified as being most important to the model + additional features
  - Sensitivity (positive or negative) to the proximity of venues