Analyzing Open Data from the City of Montreal

Joelle Pineau JPINEAU@CS.MCGILL.CA
McGill University, Montreal, CANADA

Pierre-Luc Bacon PBACON@CS.MCGILL.CA
McGill University, Montreal, CANADA

Abstract

There is a significant effort towards moving much of the data from the city of Montreal into an Open Data format. In this short paper, we report on a recent initiative to analyze this data using machine learning techniques in the context of a graduate course project. We review the approach, summarize accomplishments, and provide several recommendations for improving the impact from such efforts.

1. Introduction

Many cities worldwide have started to devote significant efforts and resources to publicly releasing data relating to their operations and situations. There is an opportunity for machine learning practitioners to use this data to answer several questions of interest for citizens, administrators, businesses, and researchers.

A course project was assigned in the context of a graduate course of Applied Machine Learning at McGill University. The stated goal of the project was to use open data from the city of Montreal's website to identify an interesting prediction question that can be tackled using machine learning methods, and solve the problem using appropriate machine learning algorithms and methodology. Previously, students had received 2 months of instructions on machine learning methods [1]. The course involved 65 students at various levels of their studies, from advanced undergraduate to Masters and PhD, 1 course in structure and 2 graduate teaching assistants. Course participants came from a diverse set of backgrounds, including computer science, electrical, mechanical and biomedical engineering, mathematics and statistics, epidemiology, neuroscience, environmental science. They worked in teams of 3 for this project.

1.1. Context and project instructions

According to instructions, participants were not restricted to using only the data from the city of Montreal website, though needed to use some of it. In particular, when appropriate, students were encouraged to incorporate data from other sources (e.g. equivalent data from other cities), or collect additional data (e.g. a new test set) to deepen their investigation.

The choice of prediction task and dataset to use was open. The goal was to pick a prediction question that is relevant and important to the citizens or administrators of the city. Particular attention was given to designing a prediction task that was well suited to the choice of dataset; and vice versa, picking the right data for tackling the chosen prediction question. The choice of algorithms and software systems was left open, including allowing use of existing machine learning toolboxes. The emphasis was on proper scientific methodology for computational analysis of urban data, rather than on the implementation of machine learning algorithms.

1.2. Characteristics of the city of Montreal dataset

The city of Montreal's Open Data resource[2] currently contains 177 datasets, organized under different themes, as listed in Table 1. Some datasets are re-listed under several themes, for example a dataset on the location and dimensions of community gardens appears under both the Environment and Housing and urban planning.

Several of these datasets include descriptive data, for example the list of municipal buildings in a particular borough, with their respective addresses, or a document describing the yearly accomplishments in terms of universal accessibility of buildings (municipal and others). In many

[1] The course syllabus:
http://www.cs.mcgill.ca/~jpineau/comp598/

[2] The data can be accessed here:
http://donnees.ville.montreal.qc.ca/

Proceedings of the 2^{nd} International Workshop on Mining Urban Data, Lille, France, 2015.

Table 1. Themes and number of datasets from the City of Montreal open data website.

Theme	Number of datasets
Organization and administration	54
Sports, leisure, culture and development	43
Infrastructures	28
Environment	27
Housing and urban planning	21
Financial resources	19
Election and referendum	17
Information management	16
Public safety and security	13
Communication and public relations	12
Material resources and services	9
Buildings and land	8
Economic development	7
Human resources	3
Legal affairs	2
Property assessment	1

respects, the data is not systematically or uniformly available: the list of municipal buildings is available for only one of the 19 boroughs of the city. The data is available in several formats (PDF, TXT, XLS, ODT, CSV, DOC, XML, KML, KMZ, GML, SHP, DXF, JSON, 3DM, ZIP), though each dataset is provided in a (small) subset of these formats.

2. Overview of projects and results

A total of 22 projects were completed, across a range of topics. Project titles are listed in Table 2. The primary challenge for most teams was to identify a dataset that contained enough data to perform a substantial machine learning analysis. This proved harder than expected, and thus several teams converged on using similar datasets from the set of 177 available. The most popular datasets pertained to the usage of the Bixi bike-sharing service, and data on the location of bicycling accidents. In some cases, participants complemented the available data with similar data from other cities, for example a project doing a comparative analysis of bicycle accidents in Montreal and New York.

A second challenge for many teams was to identify an appropriate prediction question, which was both feasible (i.e. sufficient available data) and interesting (i.e. with impact for citizens or administrators of the city). In some cases, the prediction question arose naturally out of the data, for example predicting the loan rate of library books. On the other hand, some participants were particularly creative with their choice of task. Good examples of this were found in the analysis of city images, which included a project aiming at the automatic colouring of historical images (originally taken in black&white).

Table 2. List of projects

Real estate
- Montreal Real Estate Pricing
- Prediction of Real Estate Property Prices in Montreal
- Location, Location, Location!

Transportation
- Estimating Traffic Levels in Montreal using Computer Vision and Machine Learning Techniques
- Predicting STM Bus Intervals Using Vehicles, Bicycles and Pedestrian Traffic Data
- Predicting Method of Transportation
- Biking Lane Usage Prediction
- BIXI Montreal
- Modeling imbalance in Bike Share Networks
- Predicting Bike Counts for BIXI Stations in Montreal
- Prediction Problems on Bike Accident and Usage Data in Montreal
- Prediction of Bicycle Accidents in Montreal
- Prediction of Bike Accidents, a Comparison of New York and Montreal
- Load Forecasting for Smart City with Possible Electrical Vehicle Penetration

Reconstruction/analysis of city images
- Where am I? Predicting Montreal Neighbourhoods from Google Street View Images
- Patch-Wide Classification of Historical Aerial Images of the Island of Montreal
- Reviving Old Montreal
- Object Recognition of Historical Datasets

Food safety
- Smart System for Restaurant Rating
- Predicting Severe Food Safety Violations in Toronto, Ontario

Library usage
- Predicting Montreal Library Book Loans
- Book Recommender Systems for Montreal Libraries

The choice of machine learning method to solve the chosen task was left open to the participants. In most cases, they needed to tackle the full pipeline, from feature extraction, to training the learner, to setting up a valid evaluation protocol. Many teams used common software libraries (e.g. scikit-learn (Pedregosa et al., 2011)) to assist with some portion of the work.

We now highlight a few of the projects.

2.1. Sample project: Prediction of real estate property prices in Montreal

This project aimed to predict the price of houses in Montreal. A total of 25,000 records were extracted from online listings of real estate brokers. Complementary infrastructure and geographical information for each listing was acquired from additional open data sources from the city of Montreal and Statistics Canada. Pre-processing was applied, for example removing properties with an asking price less than $10,000. Principal components analysis was used to project the feature space to a lower-dimensional space. Several machine learning algorithms were considered: linear regression, support vector regression, k-nearest neighbours, and random forest regression. Algorithms were implemented using the scikit-learn package (Pedregosa et al., 2011). The most promising results were obtained by an ensemble of k-nearest neighbour and random forest, achieving a prediction error on par with previous literature on similar datasets for other cities. In the case where the asking price of a house is included, prediction error of the selling price can be further reduced. Such a tool could be used by citizens to get a more accurate estimate of a property's market value. It may also be used by municipalities to assess property value for tax purposes. Finally, it may be used to inform economic indices.

2.2. Sample project: Biking lane usage prediction

This project aimed to predict the number of cyclists passing through different streets in Montreal on a given day. The analysis focused on ten different streets, and learned from daily counts obtained from sensors installed on the streets, over a period of dates between 2009 and 2013, with a total of 1722 records. Several features were considered, including the day of the week, weather, air quality index, price of gas, special events (festivals, football and hockey games), for a total of 47 features. This complementary data was extracted from various online sources. Several machine learning algorithms were considered: linear regression, k-nearest neighbours, boosted decision trees, and support vector regression. Prediction performance was assessed using the mean absolute error, as well as the ratio between the mean squared error for a given method and the mean squared error of a baseline (dummy) predictor. The boosted decision trees yielded the best performance. A complementary analysis of the feature impact using Lasso regression suggested that the day of the week was one of the most important features, possibly because the bicycle usage varies greatly between weekdays and weekends.

3. Discussion

In this section we discuss several opportunities and challenges that arose during the project.

3.1. Opportunities

From app design to data science. Many early open data efforts from large cities have focused on releasing descriptive data, amenable to app design, often used in the context of hackaton events. While such activities continue to be exciting and worthwhile endeavours, we believe that many communities have much to gain from also considering an open data strategy that leads to the release of urban data suitable for machine learning analysis. To meet this goal, the teams designing the open data platforms and controlling the information flow may need to acquire expertise about the goals and challenges of machine learning, in order to offer appropriate datasets. Computer scientists and statisticians have a role to play in informing these teams about the benefits that machine learning can bring to our society, and in providing convincing examples of cases where machine learning has enhanced the quality of life of citizens, and productivity of organizations.

Use of urban data to enhance transportation models. Several of the projects targeted the use of the city of Montreal data to predict various aspects of urban transportation, from the usage of the bike sharing service, to the expected timing of buses and automobiles. We observe that those datasets yielded some of the most interesting analysis because they were more extensive than other datasets, in terms of number of data points. The projects completed to date targeted specific aspects of the transportation network in isolation of others, however there is significant potential to combine such results into a coherent model of urban transportation, and eventually to use this model to evaluate different transportation strategies (e.g. adding bicycle lanes, changing bus routes, etc.)

Use of machine learning to enhance delivery of goods and services. Several of the projects attempted to use the available data to predict usage of various services, from the above-mentioned Bixi bike sharing service, to the borrowing of library books. Such analysis can be useful to make more efficient use of available municipal resources. However these cases pose particular challenges because the observed demand often depends on the availability of goods or services. So for example, one will not observe any demand for a particular book if that book was not available at the library. Similarly, it is difficult to accurately predict the real demand for the shared Bixis at a particular location once that station has no more bicycles available, and it is difficult to accurately predict demand at a new location. Some of the technical recommendations below relate to this aspect.

Use of machine learning to enhance human perception of urban data. One of the most original projects targeted the automatic re-coloration of old grey-scale images of the city. While the results so far were not fully satisfying, there is potential, as the methods improve, to use this technology to allow people to gain a new perspective on historical material. Some of the other projects relating to analysis of images have similar potential to enhance human understanding of the urban landscape, past or present.

Use of urban data as complementary data. A frequent use of the city of Montreal open data in the projects listed above was as a supplement to other more extensive datasets. An example of this are the three projects pertaining to Real estate, where a large amount of data was first retrieved from real estate brokerage websites, and then complemented (via geo-location features) with city of Montreal data on local municipal infrastructure. Additional supplementary information was also considered, from sources such as Statistics Canada (for sociodemographic indicators), the YellowPages (for location of grocery stores, medical clinics, yoga studios, etc.) and public transit authorities (for bus and subway access locations).

3.2. Teaching challenges

Methods beyond the curriculum. Several of the projects required students to tackle machine learning methods that were beyond the basic course curriculum. The lectures for the course were not designed with the final project in mind, but rather to provided good coverage of basic algorithms and methods for applied machine learning in general. Fortunately, online resources are plentiful, and most students were able to acquire the necessary material in areas pertinent to their topic. In many cases however, understanding of that material seemed to be very superficial, and more opportunity for one-on-one learning would have improved the quality of the analysis.

Managing multiple projects. One of the familiar challenges with open-topic course projects is the load it creates in terms of supervision. The instructor and teaching assistants must have the time to provide individualized advice to each project team. We observed the most intense needs during the project definition phase, with some teams requiring up to 3-4 half-hour long meetings to properly define their scope and aims.

Scope of conclusions. We observed two challenges pertaining to the interpretation of the results. First, as with any data analysis, the urge can be strong to interpret the results in ways that are not warranted by the methodology used. For example, reporting results indicating that old aerial images of the city can be classified in terms of usage type (farmland, forest, residential, water) with 80% accuracy, but failing to state that the accuracy is in fact much lower for farmland and forests, but higher for water and residential areas. Second, while quantitative results are typically the preferred metric of performance, it is often the qualitative results that speak most to the human imagination. There is a tendency to pick a few select qualitative results to “tell a story”; this can be a powerful way of showing results, but it can easily be used to mis-characterize the expected performance of a system across the full range of events.

Presentation format. Two components were used for evaluation: an in-class 3-minute spotlight talk and a written report. The spotlight talks were preferred over long talks due to the number of projects. It proved difficult to provide accurate detailed evaluations from such short presentations, and so most of the feedback was qualitative. The spotlights talks were held roughly 2 weeks before the final report was due, and thus focused more on the problem definition and methods, with few results. The final report was formatted as a research paper, max. 8 pages in length, and provided a more accurate account of the project accomplishments. In previous years, a poster session was held, instead of the spotlights and written report. This format offers more opportunity for interaction between participants. The option was not retained this year due to scheduling constraints.

3.3. Practical challenges

Language of dataset. Most of the data available for the city of Montreal is in French. Few of the resources have been translated. Even in the case of quantitative data, the lack of English-language description posed an important problem for some of the young researchers.

Design of the prediction task. When working with previously used supervised machine learning benchmarks, the target problem (i.e. *output variable*) of interest has already been identified. When working with new datasets, it can be challenging to identify the right target variable. For example in the case of the projects pertaining to transportation, it may at first seem useful to predict the number or location of bicycle accidents within the city. However these events are relatively rare, and dealing with rare events is often challenging from a statistical and algorithmic perspective (especially in small datasets). An alternative may be to predict the number of close encounters between cyclists and vehicles, which are less rare, but such data is not typically available. Alternately, predicting the flow of larger vehicles (cars, buses, trucks) may be more fruitful, since it can be reliably estimated, and can be used within a larger predictive model on urban transportation.

Lack of parallel datasets. Comparative analyses (between years, between neighbourhoods) can yield rich information. This can only be tackled if data from parallel settings is available. The well-known Boston housing dataset (Harrison & Rubinfeld, 1978) was used as a comparison for some of the projects pertaining to real estate. In general, it is useful to keep this in mind when planning for additional releases of urban open data.

3.4. Machine learning challenges

Small data. The typical ICML attendee may be tempted to believe that all the interesting tasks for machine learning deal with so-called big-data. Yet several important problems occur in the small data setting. The challenges in this case are different, possibly less computational and more statistical. There remains many opportunities to connect to the big-data community through the use of auxiliary datasets.

Sparse, incomplete, noisy datasets. As with most real-world datasets, a major problem with urban data remains the poor quality and uniformity of the data published. Often, the data is not curated by a person familiar with machine learning methods. There exists many statistical and machine learning methods to overcome problems of data quality, such as expectation maximization (Demptster et al., 1977), multiple imputation (Rubin, 1987). However the effective application of these approaches to complex datasets generally requires a good understanding of the methods (e.g. to construct a good model of imputation).

Feature coding for heterogenous data. Several projects observed that the choice of coding method for the data had a significant impact on the performance of their machine learning algorithm. For example in the bike lane usage prediction, an important feature was the day of the week. Encoding this as 7 binary features reduced the error rate by more than 5%, compared to using a single 7-valued categorical feature. Another similar effect was seen in the real estate price prediction task, where a logarithmic function was used to re-scale prices. Typically, the choice of encoding can be validated using standard methods for feature selection.

Feature selection for complex data. For some domains, the set of features that can be considered is very large, thus an important problem is in selecting the right set of features. Furthermore, it is often possible to enhance the feature set by incorporating supplementary data sources. It can be difficult to select the sufficient and necessary set of features for a given prediction task. Cross-validation methods can be used to automatically compare different feature sets. But this can be problematic in the case of small datasets where only limited data is available for validation of the feature set. An effective method in those cases is usually to use domain knowledge and expert advice to narrow down the candidate features to a manageable set (or small number of candidate sets). Another possible approach to tackle this problem is to use data from another city to predict the right set of features. Considering the case of Food Safety analysis, while Montreal has released only 750 records of food inspections (Montreal food data), San Francisco has released 10,000 records (San Francisco food data). Therefore one could optimize the choice of features using the San Francisco data and then apply the model and learn a simple prediction strategy on the Montreal data. More sophisticated methods for transfer learning are also worth investigating. Finally, it is worth pointing out that the choice of features can be key not just for building a good predictor, but also for building a good model for missing data imputation.

Choice of machine learning algorithm. There is a tendency among novice machine learning practitioners to spend significant efforts on testing several machine learning algorithms, with the belief that the choice of algorithm is the dominant factor in achieving good prediction performance. Another tendency is to assume that the most advanced methods will necessarily outperform more naive methods. In practice, several algorithms may perform equivalently, or simple methods may outperform more complicated ones, for example when there is insufficient data to properly train a complex hypothesis space, or the hyper-parameters are not properly optimized. Similar to the choice of features, algorithms can be compared using an appropriate cross-validation methodology.

Interpretability of results Methods such as linear regression, decision tree and naive bayes classifiers, are often preferred to more complex methods such as neural networks or kernel methods, in the case where interpretability of the results is necessary. In some applications, the knowledge of which features are most predictive of a particular outcome (e.g. finding which municipal amenities are best predictors of higher real estate prices) is of utmost interest. Several newer models have been proposed that combine rich hypothesis spaces with interpretability (Letham et al., To appear).

From supervised learning to decision-making So far we have been mostly concerned with *supervised* learning, where the goal of the learner is to predict a given quantity (the output) from observed variables (the input). In some cases, the goal may be to use the analysis to change a decision strategy. For example, by correctly predicting which restaurants may be found in violation of the health and safety laws, it may be possible to more efficiently deploy food safety agents. It is important to be aware of the fact that such a change in policy may result in a shift in the observed data. In the case where one wants to optimize the decision strategy, it may be more appropriate to phrase the problem under the framework of reinforcement learning (Sutton & Barto, 1998).

Off-policy learning A related case for concern arises when the data was acquired under a particular decision strategy, and the results of the analysis are used to change that decision strategy; in such case it can be difficult to accurately predict what will happen under the new decision policy. This is known as the *off-policy learning* problem in the machine learning literature (Sutton & Barto, 1998). Consider for example analyzing the usage data from Montreal's Bixi bike sharing service, then using the predictions derived from this analysis to determine which stations have lower demand, and then reducing bicycle availability at those stations. If those stations had low demand because they were already subject to reduced availability, then the further shift to reduces availability likely would not result in more satisfied customers overall.

4. Conclusion

This paper presents a recent initiative to apply machine learning techniques to analyze open data from the City of Montreal data, conducted in the context of a graduate course project. Several of the challenges and opportunities identified are commonly known in the machine learning community. Our goal in presenting this work is to illustrate how such challenges arise in the context of analyzing urban data, and in doing so, facilitate collaboration with interested parties from other communities. While the City of Montreal was not involved in the elaboration of the course project, we have since communicated results of the projects with them. We have also received inquiries from officials of other cities. There is clearly significant interest in the outcomes of such initiatives.

Acknowledgements

Much of the credit for this paper goes to the students of the Fall 2014 edition of the course *COMP-598: Applied Machine Learning*, at McGill University. The first sample project on the prediction of real estate property prices was realized by Nissan Pow, Emil Janulewicz, and Liu Liu. The second sample project on the biking lane usage prediction was realized by Robert Wenger, Haomin Zheng, and Stefan Dimitrov. Several of the issues highlighted in the discussion were extracted directly from those and other students' project reports. Additional thanks go to Angus Leigh who acted as a teaching assistant for the course, jointly with Pierre-Luc Bacon.

References

Demptster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1977.

Harrison, D. and Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 1978.

Letham, B., Rudin, C., McCormick, T., and Madigan, D. Buildling interpretable classifiers with rules using bayesian analysis. *Annals of Applied Statistics*, To appear.

Montreal food data. http://donnees.ville.montreal.qc.ca/dataset/inspection-aliments-contrevenants.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2011.

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, 1987.

San Francisco food data. https://data.sfgov.org/health-and-social- services/restaurant-scores/stya-26eb?

Sutton, Richard S. and Barto, Andrew G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.