THE SMARTER CITY: EDMONTON'S 311 EXPLORER

Carly Livingstone, MA Communication & Abhishek Mukherjee, MBA Supervisor: Dr. Olga Baysal

Carleton University

INTRODUCTION

This poster analyzes 311 Explorer, the City of Edmonton's web-based mapping tool, seeking to:

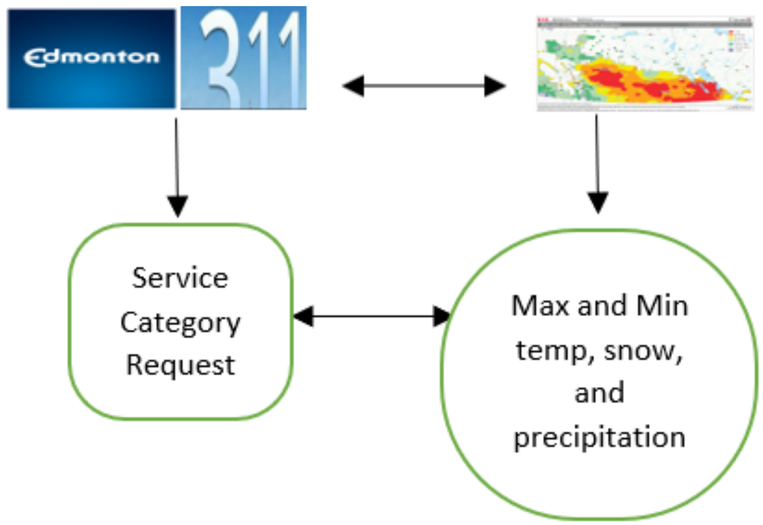
determine whether cross-referencing 311 data with historical weather data can allow Edmonton, and Canadian municipalities in general, to not only track and respond to 311 requests, but to more accurately procure, predict and allocate city resources (technologies, people and capital) more effectively and efficiently.

The use of real-time data and open data portals are increasingly used by municipalities Canada-wide to provide "smart," innovative solutions that connect city residents to information, programs and services. 311 apps and service trackers, such as Edmonton's 311 Explorer, offer data-rich tools to enable a shift from:

Responsive Cities ➡ **Predictive (Smart) Cities**

This can benefit both City officials and residents, with cities seeking to maximize efficiencies of resource use and allocation, and residents seeking faster responses to service requests in their neighbourhoods.

METHODOLOGY



DATA SETS

This project used two datasets that were cleaned, integrated and analyzed using IBM Watson Analytics, R and Tableau:

- 311 Explorer service requests dataset covering the time period of January 1, 2013 – February 1, 2018.
- Historical weather data from the Government of Canada, (observing the same time period).

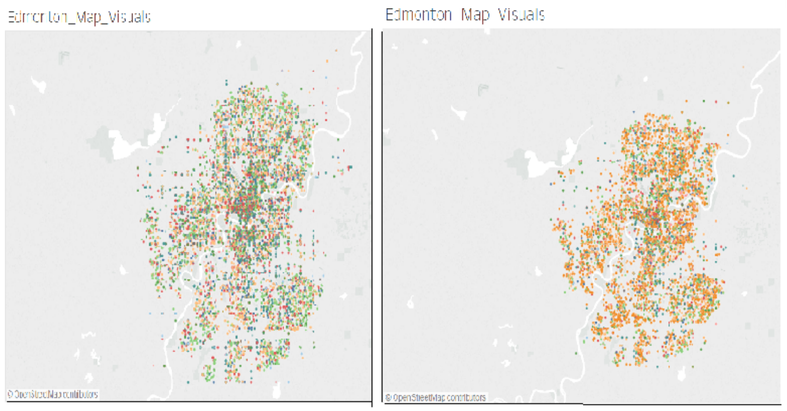


Figure 1: Image created in Tableau: Map visual representation of the integration of both data sets

SERVICE REQUEST DEMAND VS WEATHER

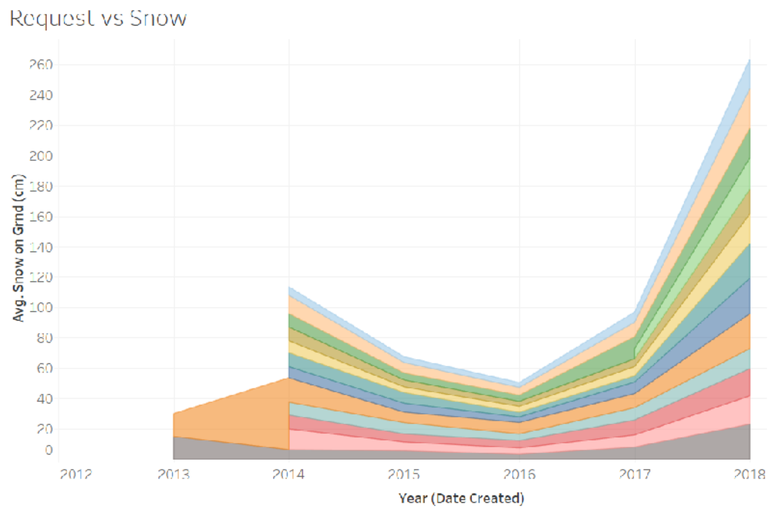


Figure 2: Image created in Tableau, representing change in service request with growing centimeters of snow

RESULTS

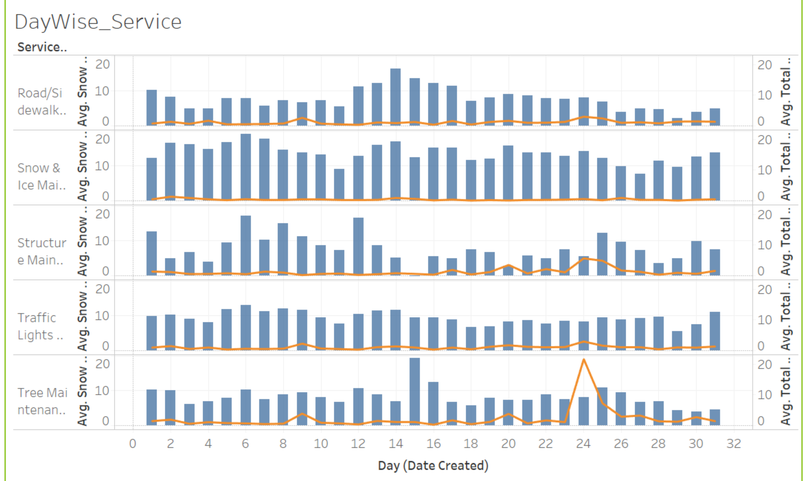


Figure 3: Image created in Tableau , representing day by day service requests in different categories

LIMITATIONS & CONCLUSIONS

Limitations

- Building predictive models based on unpredictable weather data.
- Project timeline limited to duration of course.

Conclusions

- It is possible to more effectively distribute and use city resources by observing past request data based on date.
- Using cross-referenced weather data with service requests and mapped locations, cities can better locate resources for maximum coverage of wards that require more or less assistance, and reduce response time.

REFERENCES

311 Explorer Indicators. (2018). *City of Edmonton.* Retrieved from: https://data.edmonton.ca/Indicators/311Explorer/ukww-xkmj

Station Results – Historical Data. (2018). *Government of Canada.* Retrieved from: http://climate.weather.gc.ca/historical_data/search_historic_data_stations_e.html

Minister's Message. (December, 2017). *Infrastructure Canada: Smart Cities Challenge.* Retrieved from: http://www.infrastructure.gc.ca/plan/cities-villes-eng.html

About the Service (2018). *City of Edmonton: 311 Explorer.* Retrieved from: https://www.edmonton.ca/programs_services/311-explorer.aspx

ACKNOWLEDGEMENTS: Dr. Olga Baysal, Assistant Professor, Computer Science, Carleton University

Data Science behind Canadian International Students Market

Weili Su, Zi Ye | Olga Baysal | Carleton University

Project Overview

- When people choose to study aboard, they always feel hesitate to determine where to go.
- This project is a study can help future international students make a rational decision when they choosing countries and programs.
- It is also benefit for universities since they can have a better understanding of their pricing strategy and make relevant improvements.

Research Questions

Hypothesis #1:

- What factors attract foreign students come to Canada?

Hypothesis #2:

- What is the strategy to determine university tuition fee?

Hypothesis #3:

- Is the higher tuition fee indicated a better program that has a good career prospect after graduate?

Methodologies

- Data visualization techniques to generate comparisons of related indicators between US, UK and Canada.
- Stepwise regressions to precisely track the elements that may impact tuitions (Macroeconomics factors, crime rates, university-specific indicators, immigration policies, etc.).
- Panel regression to analyze correlations between average tuitions and its respective average after-graduate incomes for archiving the last objective statement.

Models

Hypothesis #1:

$$\textbf{Tuition} = \alpha + \beta_1 \textbf{GDP} + \beta_2 \textbf{CPI} + \beta_3 \textbf{Income} + \beta_4 \textbf{Enrollment} + \beta_5 \textbf{Graduate} + \beta_6 \textbf{PR} + \varepsilon$$

Hypothesis #2:

$$\textbf{Wage} = \alpha + \beta_1 \textbf{Tuition} + \varepsilon$$

Research Variables

Hypothesis #1	Hypothesis #2	Hypothesis #3
• University Enrollments • Average Tuition Fees • Immigration Policy • Safety	• University Tuitions • GDP • CPI • Median Income Level • Crime Rate • # of Issued PR • Enrollments • Graduates	• Average Salary • Average Tuitions

Data

Source	Description
Statistic Canada	Tuition, Enrollment, Salary, etc.
CIC	New Issued PR
Government of Canada Open Data	Amount
US Department of Homeland Security	New Issued PR
UK Home Office	New Issued PR
QS World University Rankings	University Rankings
OECD	Better Life Index

Procedure

Step 1

Identify the problem

Step 2

Gather resources and data

Step 3

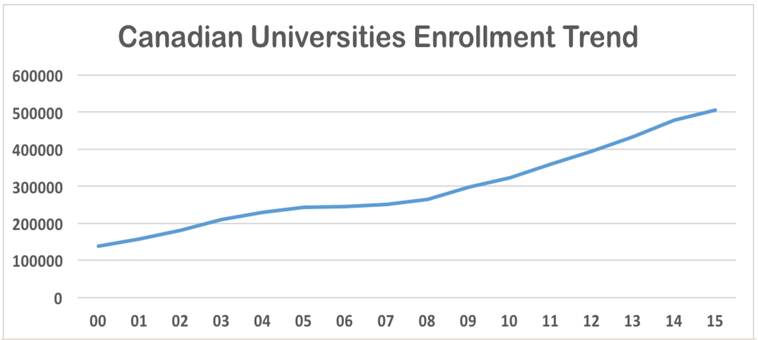
Deal with data and visualized data

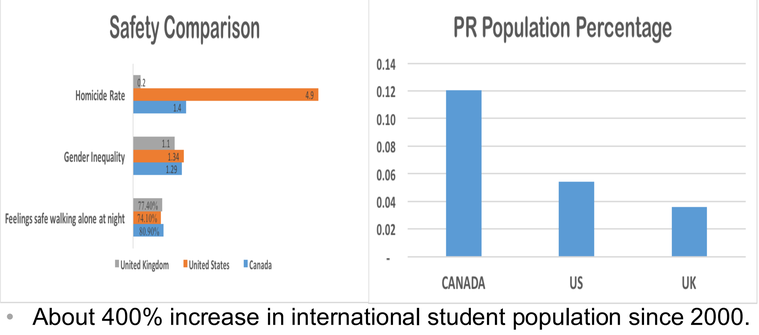
Step 4

Summarize the results from the visualized data

Observations

- Hypothesis #1:





- About 400% increase in international student population since 2000.
- Tuitions are obviously lower than US and UK.
- Safe for study and living.
- Much easier to apply for citizenships.

Results

Hypothesis #2: Undergraduate Students

- Related factors: GDP, CPI, Crime Rate & Enrollments
- Significant Variables: GDP, CPI (@10% Level)

```
. stepwise, pr(.2): regress Tuition GDP CPI Income Crime Enroll Grad PR, robust
                    begin with full model
p = 0.2939 >= 0.2000  removing Income
p = 0.6230 >= 0.2000  removing Grad
p = 0.2587 >= 0.2000  removing PR

Linear regression                               Number of obs     =         10
                                                F(4, 5)           =      75.38
                                                Prob > F          =     0.0001
                                                R-squared         =     0.9743
                                                Root MSE          =        768

------------------------------------------------------------------------------
             |               Robust
     Tuition |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         GDP |  -53.91873   22.00997    -2.45   0.058    -110.4971    2.659695
         CPI |   2143.361   719.1208     2.98   0.031     294.8022    3991.92
      Enroll |  -68.51853   40.3353     -1.70   0.150    -172.2037    35.16666
       Crime |  -6.979875   3.609617    -1.93   0.111    -16.25869    2.29894
       _cons |  -21061.08   24989.68    -0.84   0.438     -85299.1    43176.94
------------------------------------------------------------------------------
```

Hypothesis #2: Graduate Students

- Related factors: PR, CPI & Graduates
- Significant Variables: Graduates, CPI (@10% Level)

```
. stepwise, pr(.2): regress Tuition GDP CPI Income Crime Enroll Grad PR, robust
                    begin with full model
p = 0.9759 >= 0.2000  removing Crime
p = 0.3906 >= 0.2000  removing Income
p = 0.3300 >= 0.2000  removing Enroll
p = 0.4386 >= 0.2000  removing GDP

Linear regression                               Number of obs     =         10
                                                F(3, 6)           =     103.70
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9861
                                                Root MSE          =     302.47

------------------------------------------------------------------------------
             |               Robust
     Tuition |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          PR |   15.83597   8.607089     1.84   0.115    -5.224821    36.89676
         CPI |   133.6033   55.85873     2.39   0.054    -3.078085    270.2847
        Grad |   71.89199   18.51852     3.88   0.008     26.57881    117.2052
       _cons |  -21637.11   2947.818    -7.34   0.000    -28850.16   -14424.06
------------------------------------------------------------------------------
```

Hypothesis #3: Significant correlated.

```
. regress Average Tuition, robust

Linear regression                               Number of obs     =         17
                                                F(1, 15)          =      27.99
                                                Prob > F          =     0.0001
                                                R-squared         =     0.3029
                                                Root MSE          =     19882

------------------------------------------------------------------------------
             |               Robust
     Average |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Tuition |   1.923211   .3635389     5.29   0.000     1.148346    2.698076
       _cons |   46613.67   8159.682     5.71   0.000     29221.72    64005.62
------------------------------------------------------------------------------
```

Conclusion

1. Lower tuitions, safer environment and flexible immigration policies attract international students choosing Canadian Universities.
2. Determinants of tuition fee:
   Undergraduate: GDP, CPI, Crime Rate & Enrollments
   Graduate: PR, CPI & Graduates
3. Higher the tuition fee will lead to higher salaries after graduation.

Reference

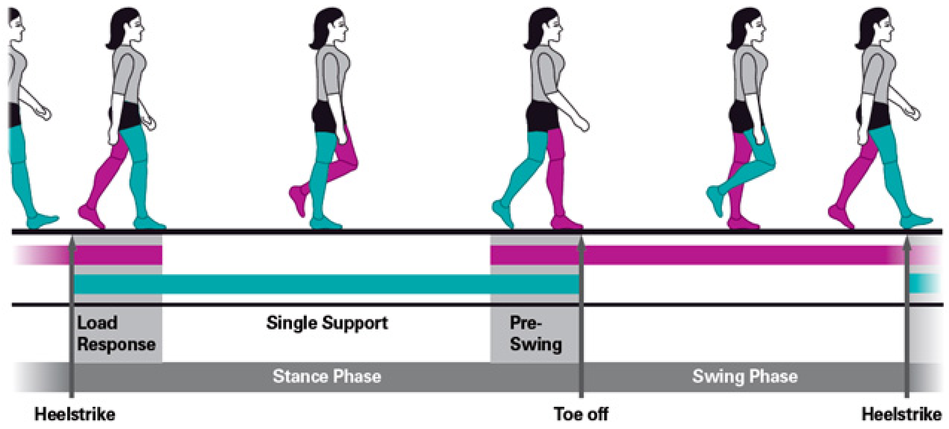
School Apply. COST OF OVERSEAS EDUCATION: US VS UK VS CANADA. 2017. https://www.schoolapply.co.in/blog/posts/2017/february/cost-of-overseas-education-us-vs-uk-vs-canada/

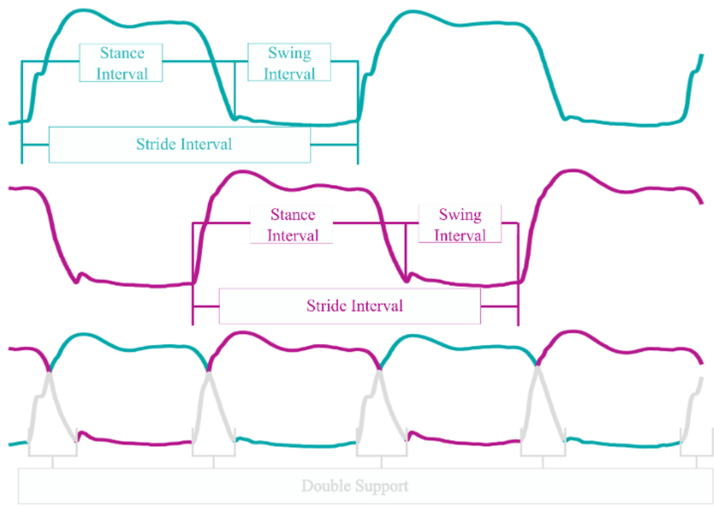
Neuro-Degenerative Disease Prediction Based on Gait Analysis Signals Acquired with Force-sensitive Resistors

Carleton UNIVERSITY
Canada's Capital University
Roger C. D. Selzler [1], Kaiqi Xue [2]
[1]Systems and Computer Engineering
[2]Sprott School of Business
Supervisor: Olga Baysal

1 Introduction

Neuro-degenerative diseases (NDD), such as Parkinson's disease (PD), Huntington's disease (HD), and Amyotrophic Lateral Sclerosis (ALS) have a significant impact on patients' daily lives, especially, on patients' gait and mobility. It is hoped that through analyzing the gait dynamic data of patients with Neuro-degenerative diseases and comparing with the date of a healthy control group, this project could reveal the relationship between the gait dynamic and neuro-degenerative diseases which could be used to predict whether a patient has neuro-degenerative disease or not.The result may be helpful to contribute to the early diagnosis of neurodegenerative diseases,thus helping with the improvement of possible therapies.



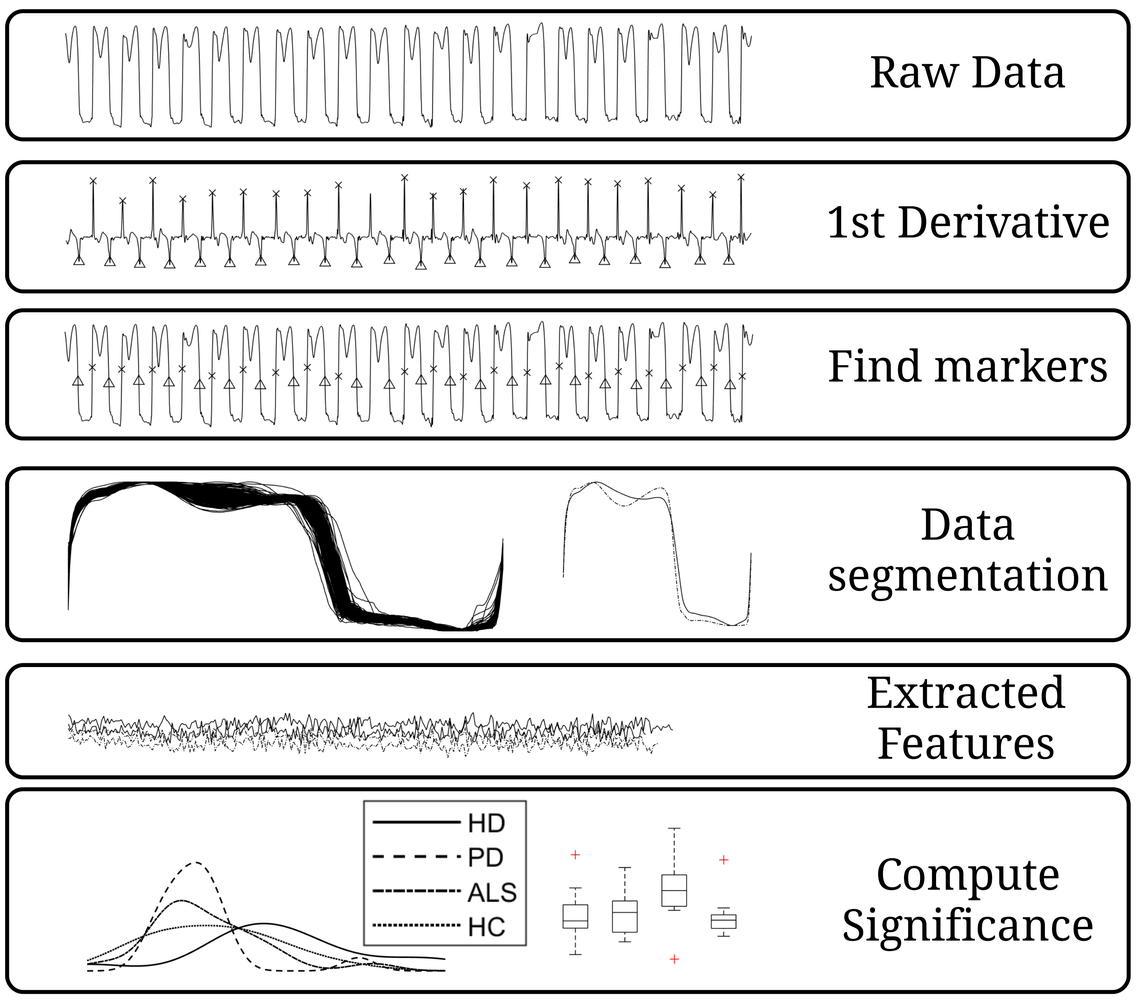


2 Methodology/Dataset

Data Source:
Physionet

Dataset:
Gait Dynamics in Neuro-Degenerative Disease

Subjects:
Healthy Control (HC) N=16
Parkinson's Disease (PD) N=15
Huntington's Disease (HD) N=20
Amyotrophic Lateral Sclerosis Disease (ALS) N=13



		Significance (p-value) ALS	HD	PD	Mean +- Std
Left stride to stride Mean (s)	HC	<0.01			1.0976 +-0.0926
	ALS		<0.01		1.4687 +-0.3619
	HD			<0.01	1.1522 +-0.1675
	PD				1.1421 +-0.1145

FS1	Stride + Stance + Swing - All labelled
FS2	Stride + Stance + Swing - Time series
FS3	Stride + Stance + Swing + FFT - Time
FS4	FFT - Time series

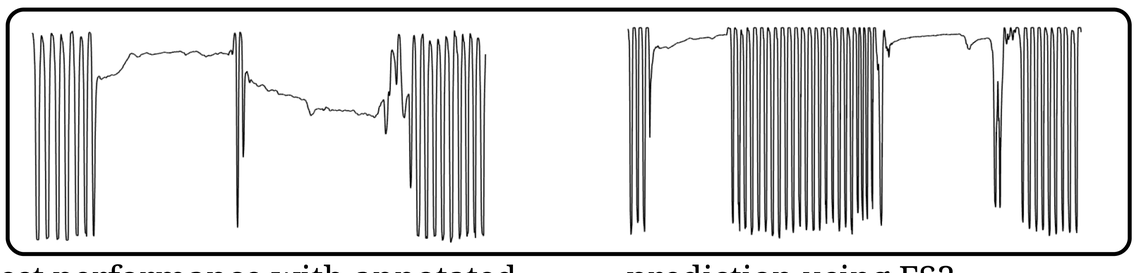
3 Results

		ALS ACC	ALS Mis	HD ACC	HD Mis	PD ACC	PD Mis	NDD ACC	NDD Mis
FS1	Best Perf	0.93 FKNN,WKNN, FT,EBAT	1 / 1	0.83 FKNN,WKNN, EBAT	2 / 4	0.86 EBAT	3 / 3	89.1 FKNN,WKNN	3 / 6
	Overall Best Perf	0.92	1 / 1	0.84	2 / 4	0.81	3 / 3	0.86	4 / 5
		EBAT							
FS2	Best Perf	0.84 EBAT	2 / 3	0.84 EBAT	2 / 4	0.84 FT	3 / 2	0.86 FT	5 / 4
	Overall Best Perf	0.84	2 / 3	0.84	2 / 4	0.83	3 / 2	0.84	5 / 5
		EBAT							
FS3	Best Perf	0.85 EBAT	3 / 2	91.7 SVML	1 / 2	0.84 SVMG	3 / 2	0.83 FT	5 / 6
	Overall Best Perf	0.85	3 / 2	0.83	2 / 4	0.81	3 / 3	0.82	6 / 6
		EBAT							
FS4	Best Perf	0.93 FKNN,WKNN	2 / 0	0.89 SVMG	1 / 3	0.77 EBAT	3 / 4	0.8 EBA	6 / 7
	Overall Best Perf	0.86	2 / 2	0.08	3 / 3	0.77	2 / 5	0.8	9 / 4
		SVML							

EBAT Ensemble Bagged Trees
FKNN Fine K-Nearest Neighbor
FT Fine Tree
SVMG Support Vector Machine - Gaussian
SVML Support Vector Machine - Linear
WKNN Weighted K-Nearest Neighbor

Accuracy (FN / FP) Algorithm

4 Conclusion



- Best performance with annotated features
  - Acc of 93.1% for ALS better than Zeng and Wang (89.66%).
  - Same Acc for HD and PD.
- Possibility of real-time processing with feature extraction on raw data
  - Acc 83.8% +- 0.046 for NDD prediction using FS2
- Increase in performance for HD by using features extracte from FFT.
  - Acc 91.7% against 83.33 in Zeng and Wang.
- Need of larger dataset to increase performance of the system.

5 References

1. Gait dynamics in neuro-degenerative disease data base. https://physionet.org/physiobank/database/gaitndd/, 2016. Accessed: 31/1/2018.
2. D. A. Wajda, R. W. Motl, Y. Moon, and J. J. Sosnoff. Stride-time variability and fall risk in persons with multiple sclerosis. *Multiple Sclerosis International*, 2016.
3. W. Zeng and C. Wang. Classification of neurodegenerative diseases using gait dynamics via deterministic learning. I*NFORMATION SCIENCES*, 317:246-258, 2015.

The H1B Visa Program and the American Dream

Changhan Kim, Master of Economics, Carleton University
Syed Tanjim Haque, Master of Business Administration, Carleton University

Introduction

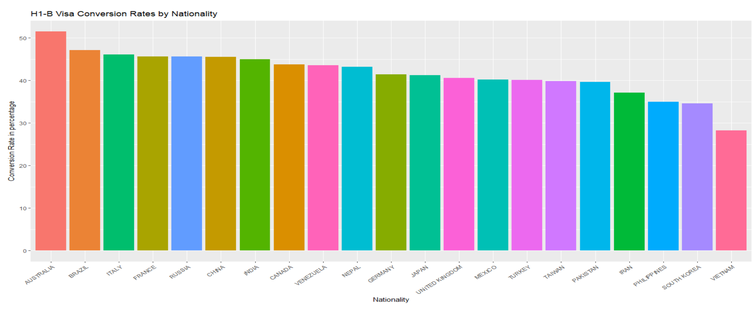
-The H1-B is an employment-based, non-immigrant visa category for temporary foreign workers in the United States.
-The motive of the research is to understand the variations in acceptance of petition of H1-B Visa across employers, locations and immigrant countries such that the results provide perspective on the petitioners in future.

Methodology

-The raw data for the H1-B Visa applicants was obtained from the United States Department of Labor website, which is generated by the Office of Foreign Labor Certification (OFLC). The dataset used contains applicants' data from Oct 1st, 2016 to Dec 31st, 2017.
-Data Pre-Processing: Deleted columns manually and reduced columns from 52 to 11. Hourly wage was converted to yearly wage by multiplying by factor 2,080 (i.e. 40 hours X 52 weeks).
-Data Analysis: Conversion rate by employers, locations, education level and nationality. Salary distribution, most denied job positions, most accepted job positions, prediction for conversion rate using the Logistic regression and the Decision Tree method.

Datasets & limitations

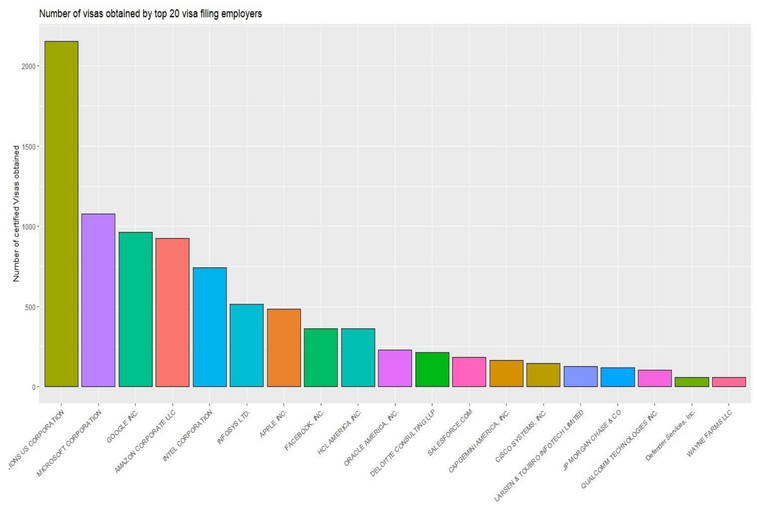
-The datasets used in the research are : case status, year, visa type, employer name, location, job title, wage offered, education and citizenship.
-Limitation: Education and nationality information is only available between Oct 2017 to Dec 2017.

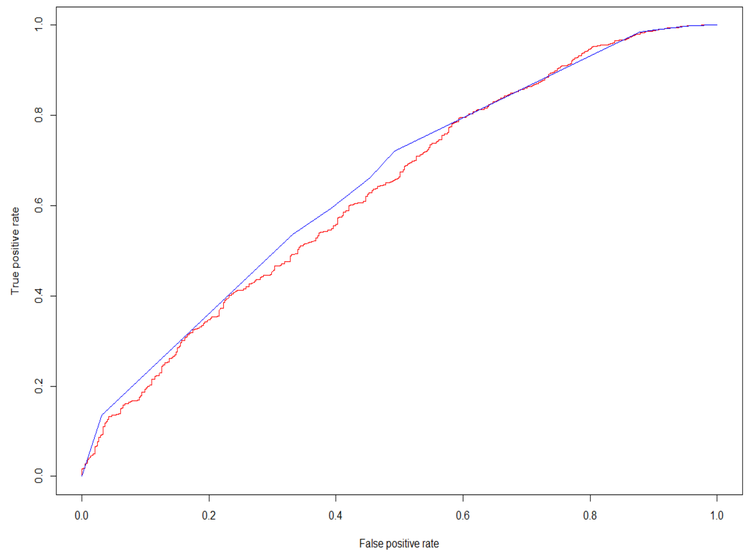


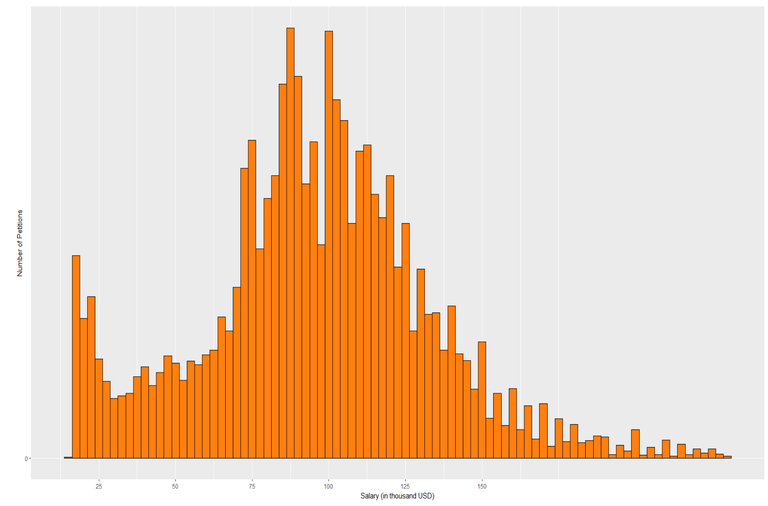
Results

Examining the data using R studio, the following results were found:

1. Number of certified H1-B Visa obtained were highest for software and electronic employers and Cognizant Tech. Corp. tops the list
2. Poultry Processing Worker has the highest rejection in terms of job profile
3. Conversion rate in terms of percentage is highest for the "Frightful 5" tech giants
4. In terms of location, California is the biggest hub for immigrants followed by Texas, NY & NJ
5. New York witnessed the highest number of Visas filed followed by College Station and San Francisco
6. Applicants from Australia has the highest conversion rate which is more than 50%.
7. Engineers and analyst have the highest conversion rate and there is a good conversion rate for Professors as well.
8. The salary distribution is a bell-shaped curve with fat tail and the salary lies between range 20k to 125k USD per annum.
9. The Area Under Curve (AUC), widely used to measure of classifier performance for binary classification tasks [1]. The AUC using the Logistic regression is 0.6588 and the score using the Decision Tree model is 0.6531.







Conclusions

-The easiest approach to obtain the H1-B Visa is to have an Australian citizenship.
-Software developers and engineers are the ones who have the highest conversion rate and it makes the conversion easier if the petitioner gets a job offer from the "Frightful 5".
-The highest number of Visa offered is seen in California, while the highest conversion rate is seen in Washington.
-The predictive power used in the prediction gives a value of 0.6588 for logistic regression and 0.6531 for decision tree model.
-It will be possible to compare before and after Trump's H1-B reform in the future.

Contact

Name: Changhan Kim and Syed Tanjim Haque
Organization: Carleton University
Advisor: Olga Baysal

References

1. [1] Shaomin Wu and Peter Flach. 2005. A scored AUC Metric for Classifier Evaluation and Selection. In Proceedings of the Second Workshop on ROC Analysis in Machine Learning. (Unpublished). DOI: https://kar.kent.ac.uk/32216/

Systematic Street View Sampling for Accurate Urban Population Estimation

François Charih[1]* and Qinrui (Michelle) Si[2]

[1] Department of Systems and Computer Engineering
[2] Sprott School of Business
Carleton University
Ottawa, Canada

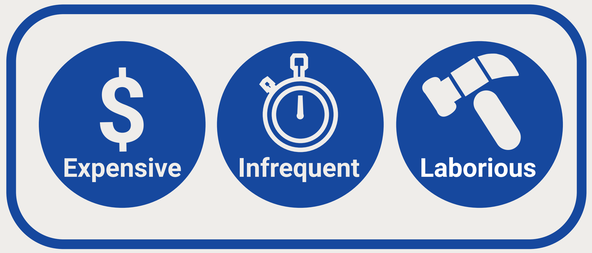
* Supervised by Prof. J. R. Green

Introduction

• Censuses provide rich information about a population and its demographic makeup.

• Policy makers and urban planners rely on population density estimates to optimize resource allocation and infrastructure development.

• Due to their cost, official censuses are only performed once every 5 years in Canada, and once every 10 years in the U.S.



Limitations of population estimation via census

• Good estimates of population count can be obtained by training deep learning models on satellite imagery [1].

• Street View imagery is expected to become increasingly abundant as self-driving cars reach the market. This provides a unique opportunity to study populations directly on the ground, at low cost and at fine-grained temporal resolutions.

• Groups have leveraged Google Street View (GSV) imagery to automate neighbourhood surveys and predict voting patterns [2, 3], but none have successfully used it to generate accurate population estimates, a logical first step in lowering census-related costs.

Reseach Objectives

Objective 1

Assemble a large, systematically sampled dataset of Street View imagery for multiple cities in the continental U.S.

Objective 2

Determine whether Street View imagery content can be leveraged to generate accurrate population estimates in urban areas.

Objective 3

Investigate the generalizability of a model trained on U.S. city data to imagery from other countries.

Assembling a Dataset of Street View Imagery

• We **collected** a large quantity of imagery from 79 U.S. cities using Systematic Street View Sampling (S³) [4], an unbiased sampling algorithm with implementations that interact with Google's APIs.

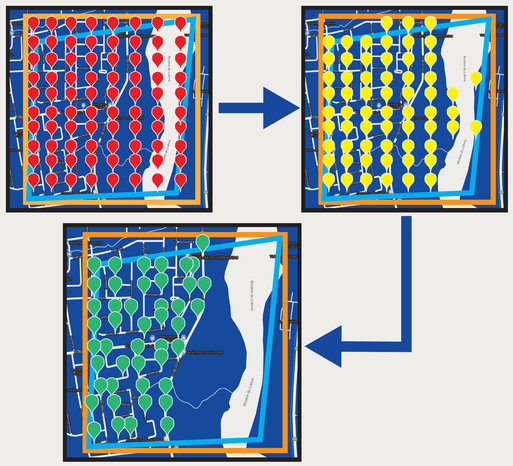


Figure 1. Systematic Street View Sampling (S³) algorithm. Candidates (red pins) are positioned along a grid defined by the region's bounding box (orange) and the sampling resolution. Coordinates outside the polygon (pale blue) or in water are removed. Remaining coordinates (yellow pins) are snapped to the nearest road. (green pins).

Sampling parameters

- Sampling resolution: **200 m**
- Number of headings per coordinate pair: **4**
- Pitch: **0°**
- Field of view: **90°**

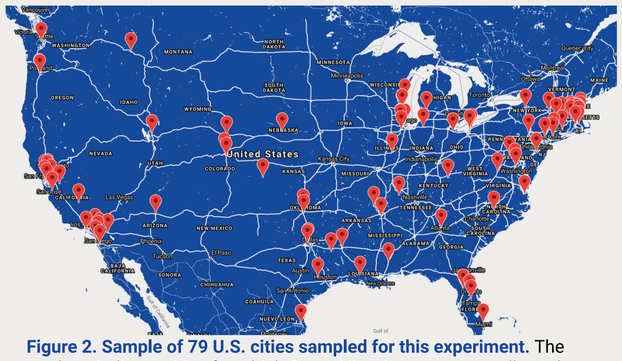


Figure 2. Sample of 79 U.S. cities sampled for this experiment. The markers indicate cities for which Street View imagery was collected.

Deep Learning for Object Counting

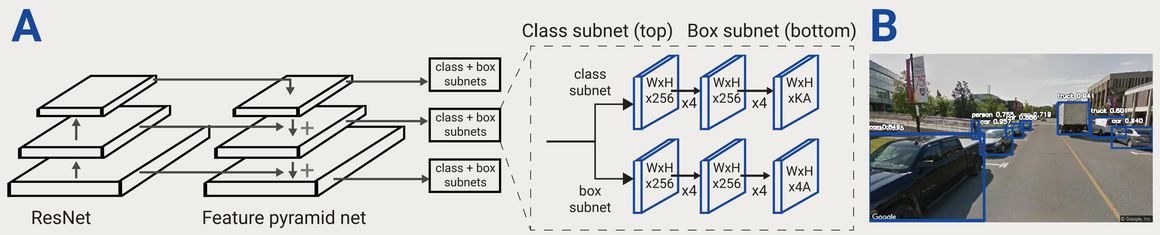


Figure 3. Fast object counting with the RetinaNet architecture. (A) Architecture of the RetinaNet convolutional neural network [5] used to count objects in Google Street View images. (B) Representative example of the application of the RetinaNet architecture on a Street View image taken on the Carleton University campus in Ottawa.

• **Deployed** a RetinaNet neural network architecture [5], implemented with Keras and pre-trained on the COCO dataset, on Compute Canada's infrastructure to make use of GPU resources.

• **Counted** 79 types of objects (people, cars, bicycles, trucks, trains, etc.) in over 425 000 images (resolution 640x380) from 79 U.S. cities.

• **Estimated** the accuracy of the object counts by crowdsourcing object counting tasks on Amazon's Mechanical Turk to estimate the accuracy of the RetinaNet network.

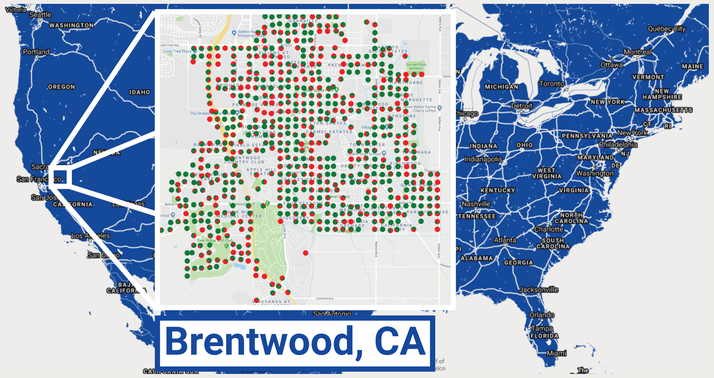


Figure 4. Street View imagery content from Brentwood, California. Circles correspond to coordinate pairs where imagery was obtained. The quadrants, corresponding to the four compementary headings of the images indicate whether objects were detected (green) or not (red).

Predictive Modelling

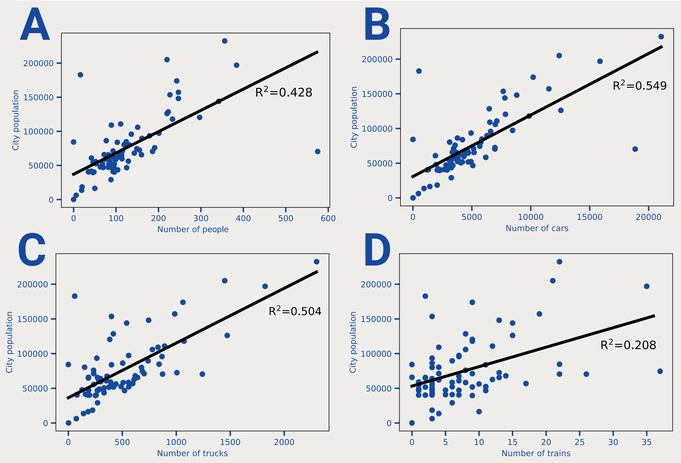


Figure 5. Correlation between the number of objects detected in Google Street View Imagery and population estimates. We computed the coefficient of determination for (A) people (B) cars (C) trucks (D) trains.

• We **tested** the performance of our predictor in leave-one-out cross-validation using U.S. Census bureau estimations over years 2010 to 2016 (weighted by the corresponding proportions of imagery) as ground truth.

Mean absolute error: **21367 ± 28568**
Minimum absolute error: **20**
Maximum absolute error: **188988**
Root mean squared error: **35675**

Conclusions and Future Work

• The accuracy of a simple model for population estimation from GSV imagery is disappointingly poor, but could be improved by including additional features and assembling a larger dataset.

• Train, using transfer learning, our object detector to count other useful objects (windows, doors, recycling bins, etc.).

• Investigate the generalizability of a model trained on U.S. imagery to other countries.

References

[1] Robinson, C., Hohman, F., & Dilkina, B. (2017). A Deep Learning Approach for Population Estimation from Satellite Imagery (Vol. 1996). Retrieved from http://arxiv.org/abs/1708.09086

[2] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences of the United States of America, 114(50), 13108–13113.

[3] Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to Audit Neighborhood Environments. American Journal of Preventive Medicine, 40(1), 94–100.

[4] Dick, K., Charih, F., Dosso, Y. S., Russel, L., & Green, J. R. (2018). Systematic Street View Sampling. In Submitted to the 15th Conference on Computer and Robot Vision 2018 (pp. 1–8). Ottawa, ON.

[5] Lin, T., Ai, F., & Doll, P. (2008). Focal Loss for Dense Object Detection.

The Slow Rise of East German Football: A Time-Series Based Performance Analysis Of The Bundesliga

Anslem Ogbunugafor (Masters in Computer Science)
Ansh Sanyal (Master of Arts, Communications & Media Studies)
Supervisor: Dr. Olga Baysal

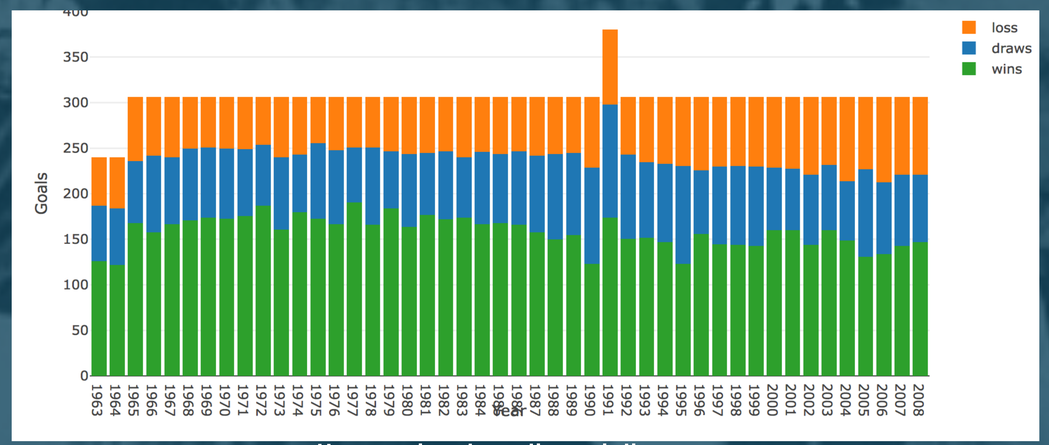
INTRODUCTION

The Bundesliga is a professional association football league in Germany and is considered to be one of the most popular leagues in the world. It "is the only European league to average more than 40,000 fans per game and is closest to 100% attendance across the course of a season"[1]. The league started in 1963 much before the reunification of Germany and comprised of teams from both West & East Germany before 1989/90. That said, West Germany teams have always been stronger in performance and competition as compared to East Germany[2]. This research study hence looks to further investigate trends and patterns around how teams have performed over past decades and in particular analyze how the performance of East Germany teams were impacted by the reunification of the country.

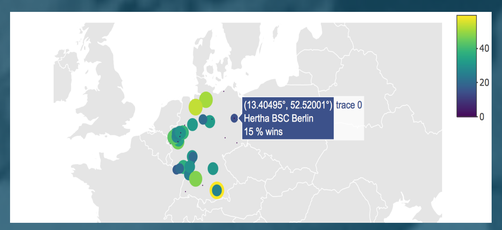
METHODOLOGY

- **Data Pre-processing:**
  - Label attribute was added for game outcome classification.
- **Data Cleanup:**
  - Nominal values were converted to numeric values and then normalized.
  - Rounds attribute removed due to low correlation at feature selection stage.
- **Training & Learning:**
  - 10-fold cross validation of the dataset was fed to models like Naïve Bayes, J48 decision tree, Ada Boost, Multilayer Perceptron.
- **Model Ranking:**
  - Models were ranked based on their F-Measure/Accuracy. Since this is a balanced dataset and are only interested in one class, "Win", the best model was AdaBoost using J48 as a classifier with an accuracy of 52%.

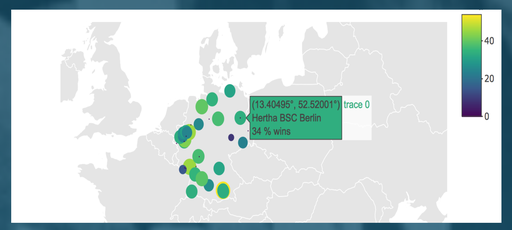
PRELIMINARY DATA ANALYSIS



Home advantage through the years



German Bundesliga, Club Competitiveness (1980)



German Bundesliga, Club Competitiveness (1990)

CONCLUSIONS

- Home advantage is important in the Bundesliga and proves to be a strong variable in terms of predictions.
- Teams originating from West Germany have been consistently more competitive compared to those in East Germany.
- Steady progress in the level of performance and competition coming out of East Germany teams.
- There is drastic rise in competitiveness amongst East Germany teams in the decade after reunifications of Germany.

IMPLICATIONS

- Football infrastructure
- Player transfer market
- Sponsorship and marketing
- Fan support
- National economy
- Socio-political impact on sports

DATASET USED

The Ergebnisse der Fussball – Bundesliga

The dataset used provides results of all games played in the Bundesliga from 1963 to 2008 and comprises of 14,018 entries and 7 variables.

TOOLS USED

- **R Studio**
- **Rapid Miner**
- **Weka**
- **Shiny**

LIMITATIONS

- **Limited set of variables in the dataset.**
- **Non-access to data around player performances and match statistics.**
- **Capability of updating dataset to reflect all years until 2017.**

REFERENCES

1. Thomas, A. (2016, December 13). Is the Bundesliga the best in the world? Retrieved January 25, 2018, from https://www.cnn.com/2016/12/13/football/bundesliga-best-league-in-the-world-five-reasons/index.html
2. Moutrie, G. (2014, October 17). West German teams still dominate the east in the Bundesliga. Retrieved March 26, 2018, from https://www.theguardian.com/news/datablog/2014/oct/17/west-german-teams-still-dominate-the-east-in-the-bundesliga
3. Dekic, D., Hothorn, T., & Zeileis, A. (2017, December 06). [Results from the first German soccer league (1963-2008).]. Raw data sourced from Package "vcd" on R.
4. Bundesliga. (n.d.). Annual Report 2017 (Rep.). from: https://s.bundesliga.com/assets/doc/1120000/1118742_original.pdf

Acknowledgement: We thank our Data Science Professor, Dr.Olga Baysal, the department of Computer Science and the School of Journalism & Communication at Carleton University for supporting and guiding this project.

Using Data Science to Develop Solutions for Fatal Road Accidents

Fahima Dawd, MBA Candidate
Supervisors: Olga Baysal & Elio Velasquez

Introduction

According to the Association for Safe International Road Travel, nearly 37,000 people die in road crashes annually in the US alone. The accumulated costs of these road incidents per year is $230.6 billion for the US government and its people.

Road traffic injuries are heading to become the fifth leading cause of death by 2030.

With road accident fatalities on the rise, it is necessary to understand the underlying factors that contribute to these accidents. In doing so, recommendations can be made to the government, drivers, and car manufacturers to improve road conditions and alleviate the impact of these factors to save lives.

Dataset & Limitations

The dataset is a complete dataset from the Fatality Analysis Reporting System in the United States with no missing values. It is harmonized and standardized as all entries are numerical vectors. These numerical vectors are codes, for which their meaning can be derived from the FARS label manual.

One limitation is that this dataset is approximately 10 years old and may not be perfectly relevant to today's accidents. However, insights can still be made as car models and roads have not changed dramatically since that time.

Methodology

Descriptive statistics were used to discover the nature of the dataset and any patterns that stood out within each column.

Furthermore, correlative analysis was conducted to better understand what spikes the number of fatalities in an accident. The predictors are sex, light condition, fatality count, injury type, and age.

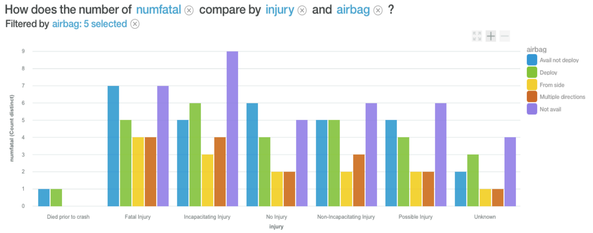


Figure 1. Number of fatalities vs. injury type and status of airbag.

modelyr	
Mean	2111.28813
Standard Err	3.68109604
Median	2001
Mode	2007
Standard Dev	937.639546
Sample Varia	879167.919
Kurtosis	66.7844444
Skewness	8.29332731
Range	8071
Minimum	1928
Maximum	9999
Sum	136982485
Count	64881
Confidence L	7.21495027

Figure 2. Descriptive Statistics for Model Year

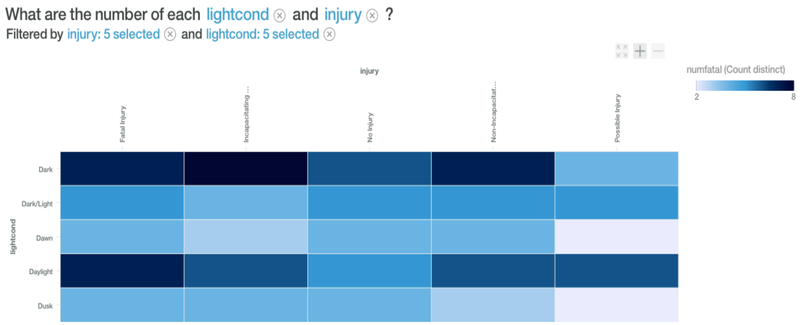


Figure 3. Severity of injury and light condition vs number of fatalities.

Results

Accidents where the airbag was not deployed had the highest number of fatalities and most serious injuries.

Descriptive statistics suggests that the car model involved in most accidents are 2007 models.

The darker the light conditions are, the likelier the injury is labeled as incapacitating and the higher the number of total fatalities.

Accidents where males are the drivers experience more in-vehicle fatality counts than females. Furthermore, the age range 18-21 has the highest number of fatalities.

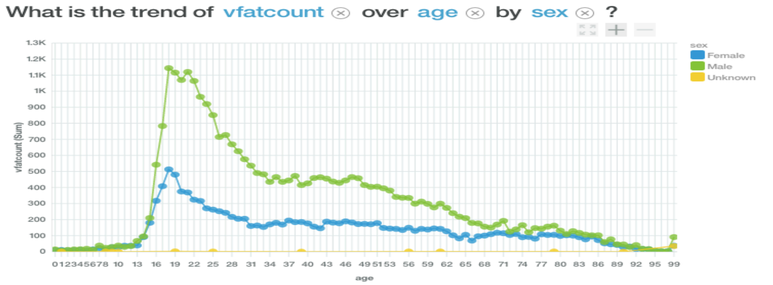


Figure 4. In-vehicle fatalities vs. age and sex of driver.

Next Steps

This analysis has confirmed the pre-assumptions previously made on airbags, light conditions and the sex and age of the driver in influencing the likelihood of a fatal road accident.

Moving forward, a regression model will be created to identify the main predictors in fatal road accidents and the predictive strength of each factor will be assessed.

Contact
Fahima Dawd
Sprott School of Business
Email:fahimadawd@cmail.carleton.ca

References

1.http://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics

Transport Collision Analytics: Characterizing Traffic Accidents in the City of Ottawa

Jesse Smith, M. Sc Geography Candidate & Steve Deery, MBA Candidate
Supervisors: Olga Baysal, Elio Velazquez

Carleton UNIVERSITY
Canada's Capital University

INTRODUCTION

Vehicle-related accidents constitute thousands of injuries and tens of fatalities in the city of Ottawa each year. Between 2012 and 2016, a total of 73,623 collisions were reported, accounting for an average of 14,724 accidents in each intermediate year.

Using collision data from 2016, our analysis seeks to develop a predictive model to uncover the factors most strongly related to collision severity and location.

The ability to understand the factors influencing collisions and their severity are extremely valuable, not only to citizens of Ottawa, but to emergency services and those responsible for traffic infrastructure.

DATASET

Retrieved from the City of Ottawa's *OpenData Catalogue*, the initial data represented 14,023 automotive collisions with 13 attributes collected and maintained by Traffic Services for the year 2016.

This dataset is of high fidelity, including no missing values and few uncategorized values (i.e. "99-other"). Character-based fields were converted into factors, and date and time into yyyy/mm/dd and HH/MM format. Geographic coordinates were supplied in modified transverse mercator but converted to latitude and longitude.

Additional attributes such as "Day of the Week", "Lunar Phase", "Holiday", "Time of Day" and "Season" were extracted from primary attributes "Date" and "Time". "Ward" was derived by matching the dataset to a base map of Ottawa wards obtained through *Statistics Canada*.

LIMITATIONS

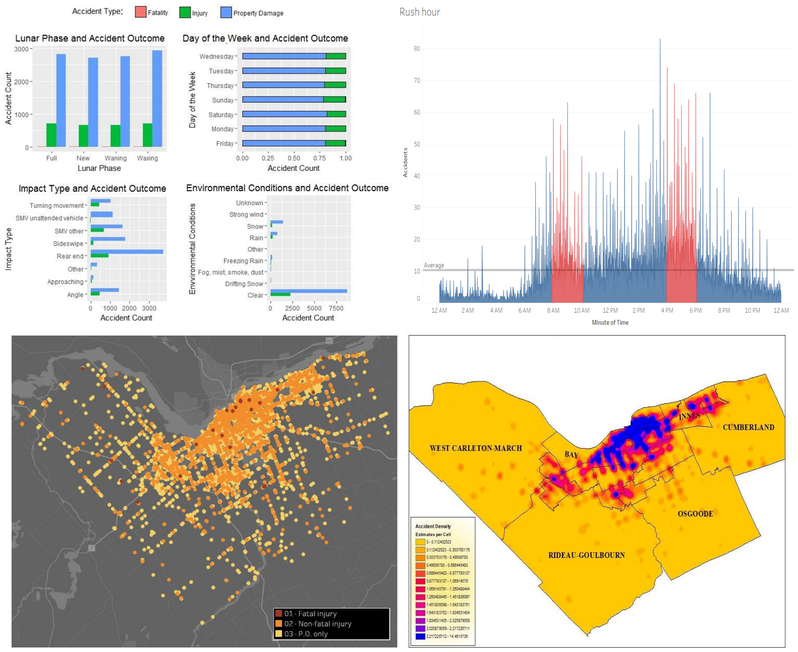
Fatal collisions make up a minority of incidents (< 0.0019). Even with over/under sampling methods (SMOTE) the initial sample size was unable to generate a 'learnable' simulated one.

This dataset does not include negative incidences (no crash). Without negative incidences, estimating the general likelihood of collision based on current conditions is impossible.

METHODOLOGY

As all provided attributes are categorical (nominal or ordinal) recursive binary splitting algorithms (rpart randomForest) were used. Multiple decision trees and random forests were produced by correlating attributes with two possible outcomes: property damage or injury.

SMOTE was used to adjust the balance of injuries (minority class) to approach the number of cases of property damage (majority class).



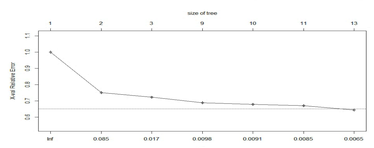
Exploratory analysis of data support our initial "common sense" assumptions.

Top: Total collisions increase during daytime, specifically during morning and afternoon rush hour (red) (right). Sample visualizations that illustrate the class imbalance problem (left). Bottom: Density map of collisions by city Ward (right) and a point map indicating fatalities cluster in the inner city (left).

RESULTS

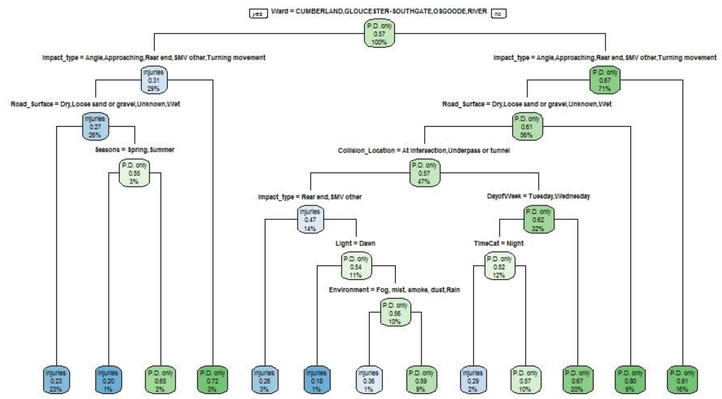
Preliminary results suggest "Ward" to be the strongest predictor of collision severity. "Impact type" and "day of the week" are the next most important attributes in predicting accident outcome.

	rcart	Random Forest
Precision	0.5475761	0.9496628
Recall	0.7331293	0.9913262
Accuracy	0.7206784	0.9748619

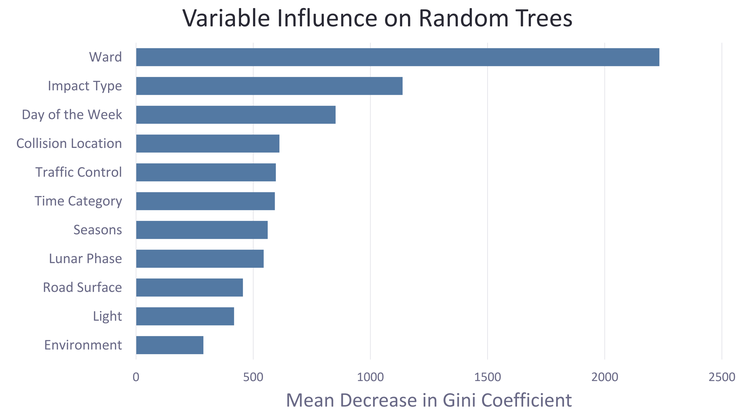


Confusion matrix of classification tree and Random Forest performance (left), **complexity parameter plot** (right).

RESULTS



Sample decision tree created with rpart. The size/complexity of the tree is optimized by selecting the complexity parameter with the minimal cross-validated error.



Ranking variable importance with Random Forest. "Ward" is the most influential and doubles the explanatory power of the next most influential variable. This suggests that ward-specific modelling should be explored as a next step.

FUTURE DIRECTIONS

This analysis developed a functional model to classify collisions based on their attributes.

The next step is to produce a "master" decision tree used to predict accident type within individual wards. Wards that score below a threshold metric will have their own trees built to evaluate spatial variability.

Results will be further validated using a separate model created with ordinal logistic regression.

Lending Club's Issued Loans Analysis

Xingyu Pan (Economics)
Ruijie Zhang (MBA)
Advisor: Olga Baysal & Elio Velazquez

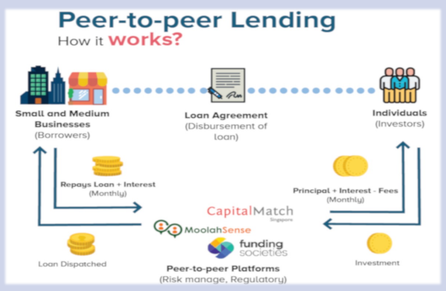
Introduction

Lending Club is the world's largest online marketplace connecting borrowers and investors. It is transforming the banking system to make credit more affordable and investing more rewarding. Peer-to-peer investment has become a much more mainstream part of financial services in recent years. But there is a high rate of defaults hit P2P lending sector.

LendingClub

How P2P lending works?

- If customers have interest in a loan, they can complete a simple application at the Lending Club website.
- Lending club evaluate the grade and determine appropriate interest rate and the amount of loans.



Research Problems

Aim:

Analyzing the features of loan lend by the Lending Club Company in order to help the company and individual investors to avoid default risks and predict the default rate.

Research Question:

- RQ1: What's the basic loan status of loan issued by Lending club in the third quarter in 2017.
- RQ2: Is the default rate predictable? How to help Lending Club issue more quality loans?

1) What kind of borrowers pay monthly repayment on time?
2) Is there profit if we lend money to low grade borrowers?
3) What do borrowers do with the money they borrow?
4) Is there a strong relationship between the loan amount and installment, the grade and loan term, or the monthly payment and the principal amount?
5) Is the results of prediction reliable?

Methodology

Data Visualization

- Observe characteristics of borrowers.
- Analyze their repayment behavior and their purposes of borrowing a loan.

K-Nearest Neighbor

- A classification and a regression model. An object is classified by a majority vote of its neighbors.
- Evaluate the performance of the model by using the prediction against the actual class of loans in the test data set.

Decision Tree Algorithm

- Use decision tree as a decision analysis to identify the quality of borrowers.
- To predict whether a loan will be paid off in full or the loan will be charged off and possibly go into default.

Logistic Regression Model

- Find the attributes that statistically significant related to the probability of default.

Data Preparation
- Missing values
- Converting few columns as factors
- Train Data/ Test Data

Modeling
- K-Nearest Neighbor
- Decision Tree
- Logistic regression

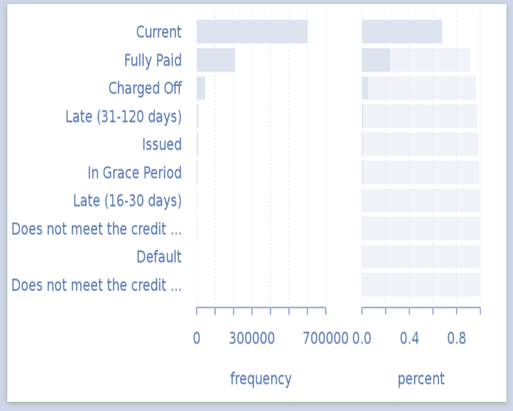
- Visualization
- Result Analysis
- Model Comparison

Results

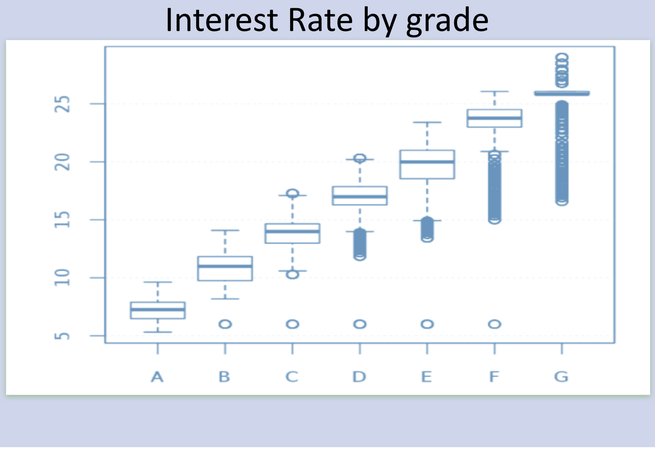
Logistic Regression Results

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.208e+00  5.477e-01  -5.858 4.69e-09 ***
loan_amnt   -1.587e-05  2.566e-06  -6.184 6.25e-10 ***
annual_inc   1.368e-06  5.121e-07   2.672  0.00755 **
dti          2.249e-03  1.683e-03   1.336  0.18152
fico         1.016e-02  7.925e-04  12.815  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Loan Status



Interest Rate by grade



Model Comparison

- We compare the performance of these models by measuring the **ROC area** and using **confusion matrix**, to see which model has highest accuracy value.
- Or using other various values in fit details, like: **RMSE, R square value.**

Datasets

The dataset contains the newest information about loan issued in the third quarter in 2017.
There are around 151 features, and ach feature contains 120,000 data.
Keep 30 valuable columns, and separate the whole data to train data and test data.

Conclusion

- The default rate depends on the given attributes. For instance, low grades loan are more likely to be default.
- Compare the performance model and find that the data mining algorithms are effective on analyzing loans.
- Helping Lending Clubs issue high credibility loans. Charging proper interest rate for each loan.

Acknowledgments

- This project was supported by the Lending Club official data.
- Special thanks to professor Olga Baysal who has provided valuable suggestions throughout the project.

[1] Jin, Y. and Zhu, Y. (2017). A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending. IEEE, [online].
[2] li, l., shrestha, s. and Hu, G. (2017). Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques. IEEE, [Journal].
[3] Verver, J. (2017). How analytics give governments greater oversight of risks and controls in P2P processes. [online] ACL.

MODELLING AIRBNB GROWTH AND OPPORTUNITY IN TORONTO

Jessica Guillemette
Masters of Arts: Economics
Professor Elio Velazquez (DATA5000)

Carleton UNIVERSITY

Tahira Ghani
Masters of Computer Science
Professor Elio Velazquez (DATA5000)

INTRODUCTION

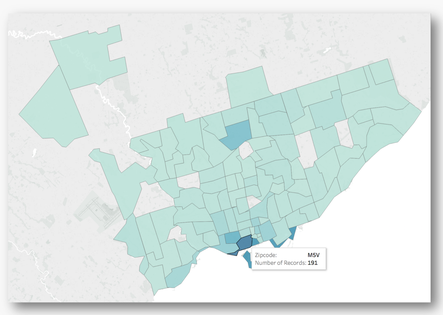
- Consumer migration towards a **"sharing economy"**[1] whereby short term access to certain assets can be acquired for a cheaper rate
- Examples of such services include ridesharing companies (Uber, Lyft, etc.) and home sharing companies (Airbnb)
- Airbnb, with its rapid growth since launching in 2008, has **revolutionized** the way consumers are **vacationing**
  - This trend has massively impacted the tourism industry, allowing a wider range of people the ability to travel with affordable lodging
- CRITIC: It has **reduced supply of long term rentals** for individuals that need an apartment to live in[2]

OBJECTIVES

- Focusing on **Toronto, Ontario** as it is one of the most toured cities in Canada
- Analyze the considerations of both Airbnb hosts and clients
  - Expectations of **annual income** from Airbnb
  - **Property types** (house vs. apartments) for investments
- Main purpose(s) of our analysis is to:
  - Analyze which factors are relevant in posting success
  - **Predict success** of an Airbnb posting, i.e. have a high rating (4 or 5 stars)
  - Forecast prime neighborhoods in Toronto for property investment through **calculating percent yielded** by hosting it on Airbnb

DATASET

- Acquired publicly available data from open source community, **InsideAirbnb**[3], which scraped data from Airbnb
  - Used Toronto data, last compiled in June 2017
- Dataset includes:
  - Detailed and summarized listings
  - Detailed and summarized reviews
  - List of neighborhoods in the city
- Compiled a small, aggregate dataset on **real estate properties** in Toronto
  - Disclaimer: All information and rights belong to Royal LePage[4]. Data collected was limited and for personal use.



METHODOLOGY

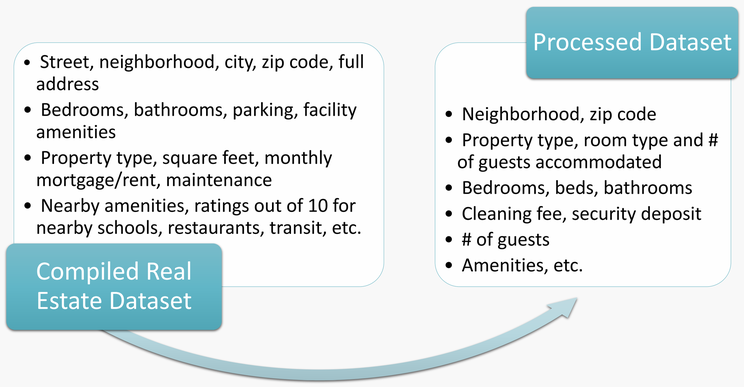
PART 1

- It is important to note that the following process flow was followed for two different types of predictions: **SUCCESS** and **PRICE** prediction, hence two models were chosen based on the highest accuracy and evaluation metrics

Preprocessing	Feature Elimination	Model Selection & Testing
• Numerical categorization of features • Matching data types to information • Outlier removal • List-wise deletion	• Correlation threshold • Principal Component Analysis	• Random forest • Bayesian additive regression trees (BART) • Boosted regression trees • Logit (logistic regression)

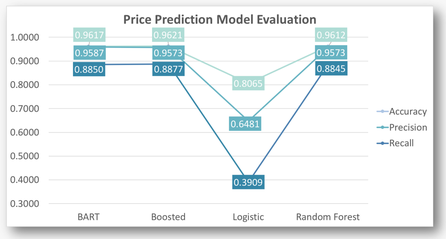
PART 2

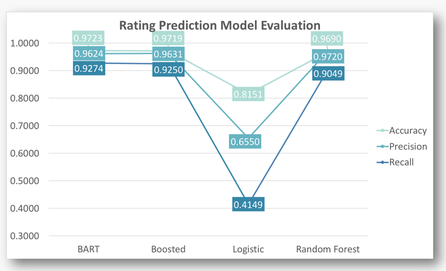
- After model selection, the real estate dataset was formatted to match the selected feature information and processed so that **each property was assigned** a Airbnb **rental price** and **potential success rating**
- The assigned estimates and ratings were then used to calculate a **percent yield** or **annual income** of a potential Airbnb posting



RESULTS

- Results were acquired using **cross validation**
- Models were evaluated on **accuracy**, **precision**, and **recall**
- The charts below show comparison between the models that were tested
- It is important to note that the ratings and pricing were divided into four equal parts based on their **quartiles**
- Part 2 Results are **to be determined** (processing in progress)
- The best model for rating prediction is **BART** and for price prediction is **Boosted Regression Trees**





CONCLUSION

- The model chosen to be used in Part 2 is the best model for price prediction, **Boosted Regression Trees**
- The accuracies of the different models (excluding logistic) were similar – the recall and precision were used to help choose the best model
- Eagerly working on acquiring Part 2 results
- Future work can include adding **more dimensions** of data to see the effect on prediction and key investment areas
  - Example: crime data of a neighborhood

airbnb

ACKNOWLEDGEMENTS

We would like to thank **Professors Elio Velazquez** and **Olga Baysal** for their guidance and the opportunity to present our work. We would also like to thank **InsideAirbnb** for allowing us to use their collected data on Airbnb.

REFERENCES

[1] Oxford Dictionaries. *Definition of sharing economy.* 2018. URL: https://en.oxforddictionaries.com/definition/sharing_economy

[2] Giovanni Quattrone et al. "Who Benefits from the 'Sharing' Economy of Airbnb?" In: *Proceedings of the 25th International Conference on World Wide Web.* WWW '16. International World Wide Web Conferences Steering Committee, 2016, pp. 1385-1394. ISBN: 978-1-4503-4143-1.

[3] InsideAirbnb. URL: https://insideairbnb.com/get-the-data.html

[4] Royal LePage. URL: https://royallepage.ca

ELECTIONS

POLITICAL DONATIONAL ANALYSIS

Analyzing Factors That Influence Political Donations to Federal Candidates in Canada

OBJECTIVES

We intend to gain insight on what factors political parties need to focus on to maximize their donor contributions.

WHY IMPORTANT

Election campaigns are expensive, so how to efficiently donations are solicited from constituents becomes a significant issue. Soliciting donations more effectively can not only result in a higher amount of overall donations, but also reduce the overall expense of soliciting those donations. These additional resources can then be allocated to other functions that help make a party

INTENDED OUTCOMES

Through this analysis, we hope to produce a profile of the types of donors that have historically donated to each party, and thus inform whom they should focus canvassing efforts on.

DATA

Data used in this analysis includes: Elections Canada 2004-2017 political donations dataset, income level datasets of metropolitan cities in Canada from Statscan, and additional population density dataset, crime rates dataset, and educational level datasets.

ASSUMPTIONS

Using the average amount of a donation to represent the efficiency of soliciting donations; Assuming the relationship between various factors and the average amount of a donation is linear

METHODOLOGY

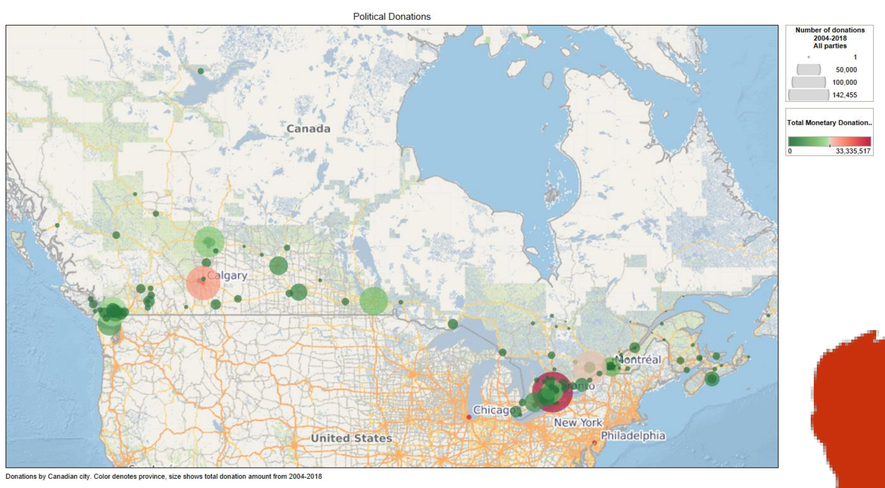
The methodology involves linear regression analysis on the relationships between various factors, such as income level, population density, and crime rates, and the average amount of a donation within each city. This analysis will include significance testing of each factor introduced.

RESULTS AND FINDINGS

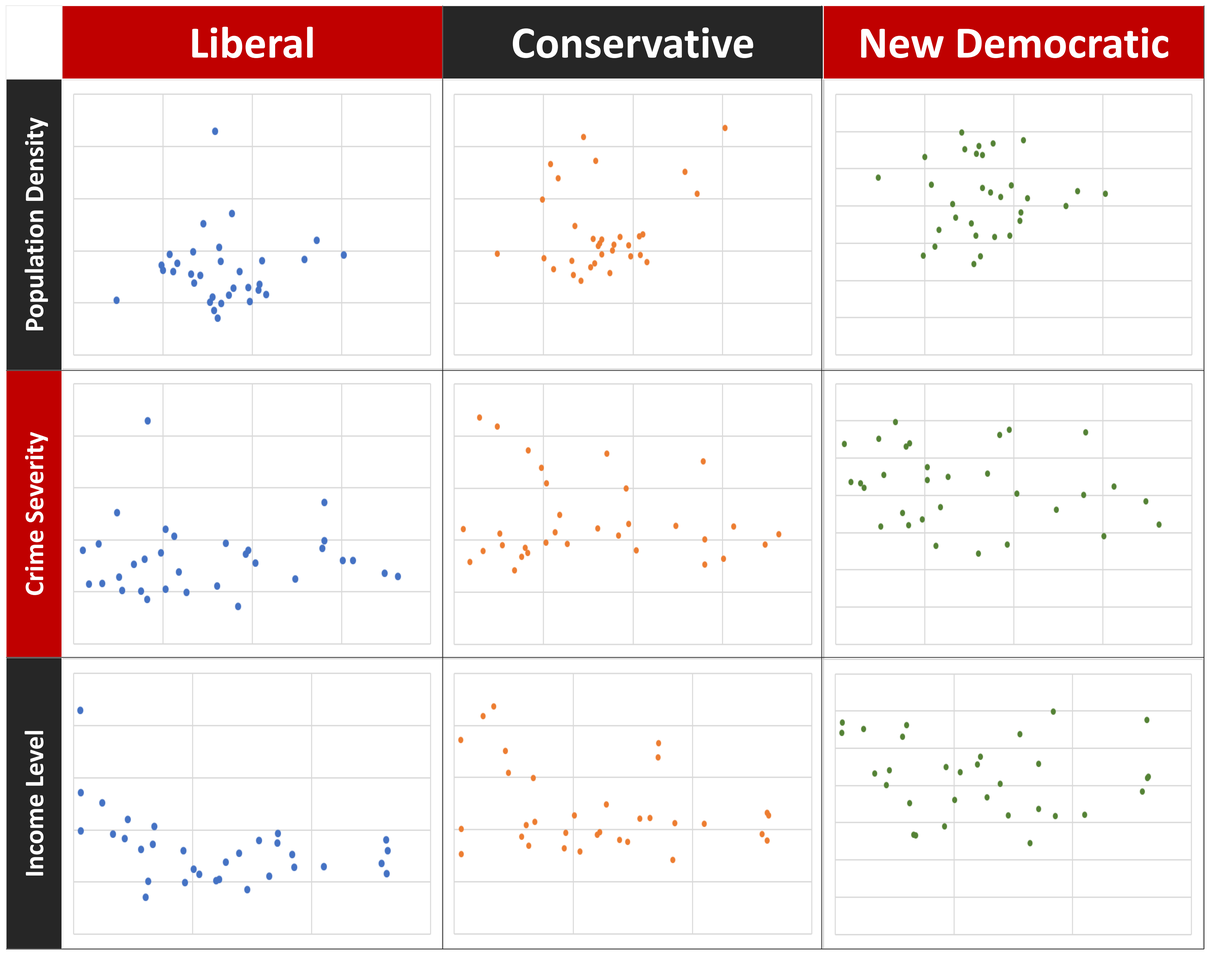
Using Excel to run multiple regressions on the data, we had got beta coefficients of each model. Based on the results of Liberal Party, Conservative Party, and New Democratic Party in 2015, 2014, and 2013, we found that none of the betas was significant and the R square of each model was very low. This implied that the average amount of each donation was not predictable by the chosen factors. The chart on the right shows the correlation between the average amount of each donation and various factors is weak.

LIMITATIONS

Linear regression shows a theoretical relationship between various factors and the average amount of a donation, but this relationship may not be linear. In addition, the models cannot reveal causality, which means canvassing efforts that focus on any one factor may not necessarily increase the average amount of a donation.



Donations by Canadian city. Color denotes province, size shows total donation amount from 2004-2018



Luis Padilla-Ramirez
Junjie Mai
March 27

The Challenges of Integrating Student Diversity in Canada

Hanru Chen – hanruchen@cmail.carleton.ca, Yao Song – yaosong@cmail.carleton.ca;
Supervisor: Olga Baysal

INTRODUCTION

As an international educational center, Canada attracts millions of students globally every year. Since the trend of globalization seems irresistible, the increasing cultural exchanges will effectively improve mutual understanding and friendship between Canada and other countries. Noted the increasing amount of international students, how to balance this diverse group and local students well is a significant challenge. The aim of this study is to analyze the market trend and factors for possible market changes in the Canadian international student market, which is useful for the market development in the future.

RESEARCH QUESTIONS

To figure out the main problem stated above, we should specifically think about some issues in the report as follows:

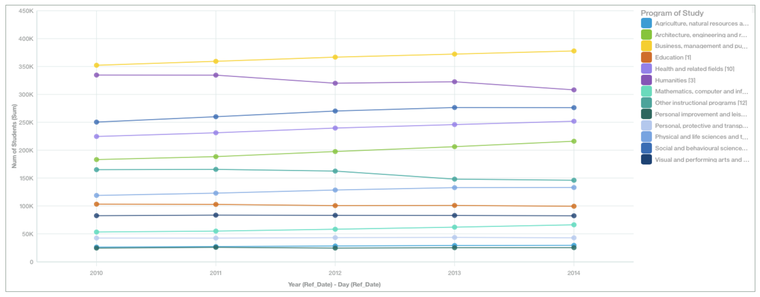
- How international students vary studying in Canada by province, program, etc.
- The preferences in the international student market in Canada with comparison of nationality, sex, level of study and so on.
- The relationship between program tuition and graduate earnings.
- The current situation that international students apply for permanent residents in Canada after graduation.
- The popularity of Canada as an ideal abroad study country, regarding political, economic, social and legal aspects.

DATASET

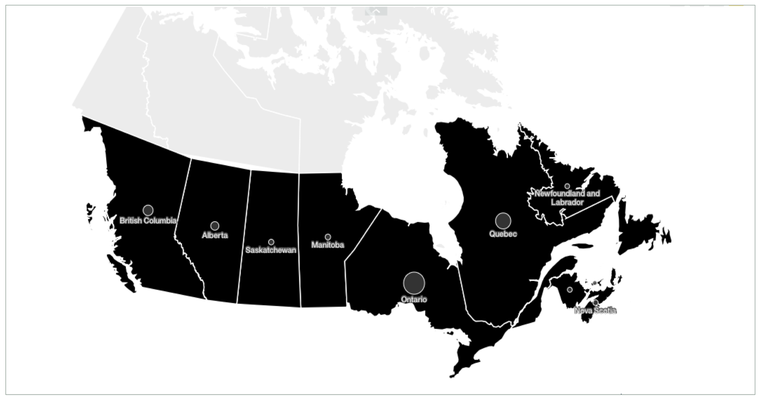
The dataset includes the number of international students studying in Canada sorted by province, program, nationality, tuition fee vs graduate earning figures, the number of international students choosing to remain in Canada as permanent residents after graduation and etc. respectively during the period between 2010 and 2014.

METHODS

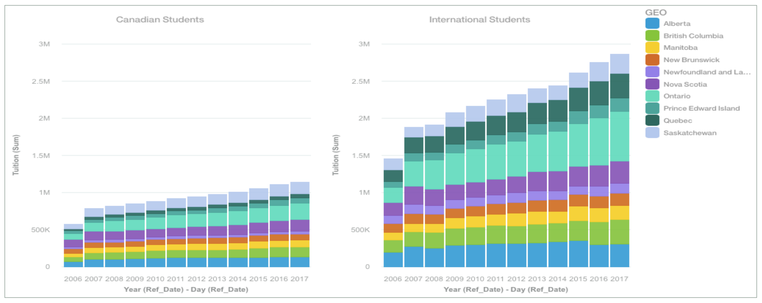
- Program Diversity



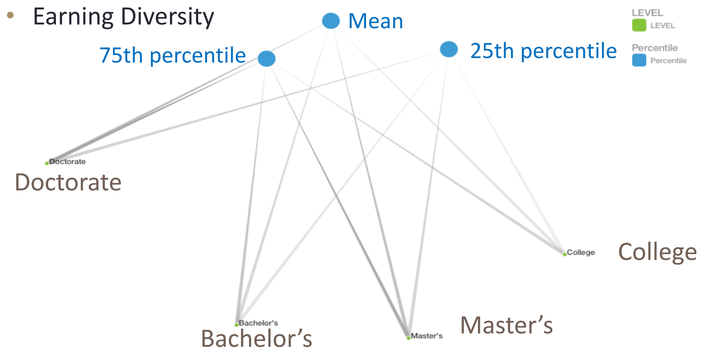
- Province Diversity



- Tuition fees Diversity



- Earning Diversity



IMPLICATION

- This predication can be used for Canadian Universities to know better about their student diversion in the future.
- In term of education policy, the results suggest further research on the ways in which institutional characteristics influence the development of a diverse student body, since these are clearly important. They also suggest that schools located in regions with ethnically and racially homogeneous populations will find the creation of a diverse student body more challenging.

RESULTS

Canadian student:

- Ontario 55149.00
- Quebec 37310.54
- British Columbia 18384.88
- Alberta 13313.72
- Manitoba; Nova Scotia; Saskatchewan
  - Program A* 6062.29
  - Program B** 1541.13
- New Brunswick; Newfoundland and Labrador
  - Program A* 3251.74
  - Program B** 655.45
- Prince Edward Island 479.68
- Territories 248.26

International student:

- Ontario 6028.11
- Quebec 2693.68
- British Columbia 2622.05
- Alberta 1148.12
- Manitoba; Nova Scotia; Saskatchewan
  - Program A* 642.20
  - Program B** 107.93
- New Brunswick; Newfoundland and Labrador
  - Program A* 307.46
  - Program B** 105.05
- Prince Edward Island ***
- Territories 0.05

Not reported:

- Ontario 19.29
- Quebec 0.00
- British Columbia 0.14
- Alberta 4.38
- Manitoba; Nova Scotia; Saskatchewan 56.39
- New Brunswick; Newfoundland and Labrador 2.51
- Prince Edward Island 61.54***
- Territories 10.15

CONCLUSIONS

- Many factors may affect student diversity at all levels of study, but local demographic characteristic is likely to be a critical one.
- As a popular international studying hub, Canada will still attracts more international students in the future. However, the differences among different provinces and different programs are not going to be eliminated.
- Most graduated international students applying for permanent residence are from developing countries.

Notes:

- Program A*: Architecture, engineering and related technologies; Other instructional programs; Physical and life sciences and technologies; Business, management and public administration; Health and related fields; Humanities; Social and behavioural sciences and law;
- Program B**: Agriculture, natural resources and conservation; Personal improvement and leisure; Education; Visual and performing arts and communications technologies; Mathematics, computer and information sciences; Personal, protective and transportation services;
- The total number of international students and not reported student in Prince Edward Island is 61.54.

Reference:
[1] CIBC, Deloitte Canada, Tableau. (2018) Data vizart: The Student Challenge 2018. http://www.thestudentchallenge.ca
[2] Rachel S. Franklin. (2013) The Roles of Population, Place, and Institution in Student Diversity in American Higher Education. https://onlinelibrary.wiley.com/doi/abs/10.1111/grow.12001

Success and Failure in Crowdfunding: A Look into the Data

Ashiqul Haq Chowdhury, Master of Economics; Navid Hossain, Master of Engineering: Electrical and Computer
Supervisor: Olga Baysal

Introduction

This project discusses about analyzing data on projects seeking funds from Crowdfunding sources (Kickstarter) and looks for patterns in their success or failure in raising this fund based on their product category, fund goal, backers etc.

The project looks to primarily address the following questions:

- Is any particular product category more successful in general than others? This may reflect a pattern in funder's behavior i.e. they are more interested in funding some product categories over others.
- Does average pledge (total fund raised/no. of funders) say anything about the success of project?
- Evaluate how the model we propose succeeds in predicting the success and failure of the projects accurately

Methodology

In this project, we are using logistic Regression as our evaluation method for the dataset. Logistic Regression method is particularly useful when we require to estimate the probability of an event occurs for a randomly selected observation vs the probability that it does not occur.. In our case, we need to find a pattern in success and failures of different projects, whether the success rate varies for different categories. Therefore, it seems more appropriate to us to use logistic regression method. We use the following specification:

$$P(state) = \frac{e^V}{1 + eV}$$

Where, State = 1 for Success & 0 for Failure of a Project

$V = \beta_{i1}(Category) + \beta_2(Project\ Goal) + \beta_3(Pledge/person)$

i = no. of main categories of (15)

We used 80 percent of our sample to train the model and test the results on the rest 20% of the sample to evaluate the model. We also use the same specification to run a regression decision tree and use the results for a comparison with the logistic regression model.

Datasets and Limitations

For the project, a dataset containing name, category, country of the project, currency the fund was raised in, project launch and deadline date, funding goal, pledged amount, no. of backers/funders, and state (successful/failed/cancelled/live) of about 300,000 Kickstarter projects will be studied. This was collected from Kaggle.

Although there is a large number of data in the dataset, there is a lack of variety in the data. Another alternate dataset is required to analyze and evaluate some of the parameters of the dataset properly.

Results

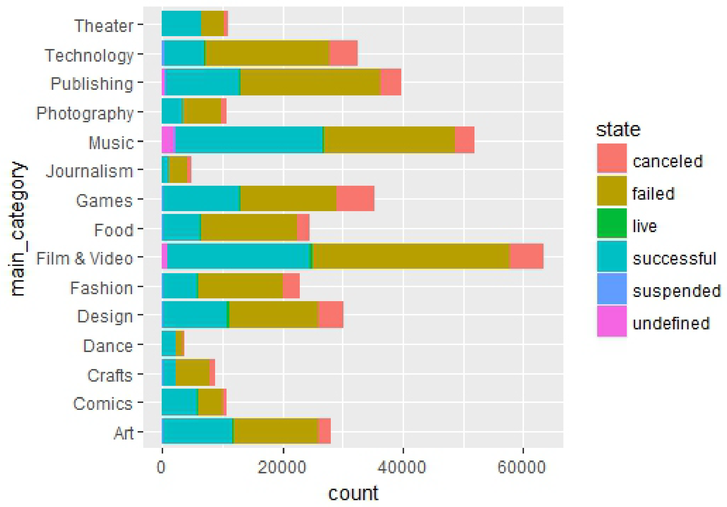


Figure: State of Projects by Category

The projects in the dataset are labeled as one of the six states: failed, successful, canceled, live, suspended, undefined. We focused our analysis on the projects which were labeled 'failed' or 'successful'. The number of observation for this study was 331,675 (197,719 failed, 133956 successful)

		Coefficients
	Top 3 Categories	
1	Theater	0.7025
2	Dance	0.6999
3	Comics	0.5245
	Bottom 3 Categories	
1	Technology	- 0.7921
2	Crafts	- 0.7098
3	Fashion	- 0.7025
	Other Variables	
	Goal	$- 2.017X10^{-5}$
	Pledge/Person	0.00575

Table-1: Results from Logistic Regression

The table represents the coefficients of the logistic regression of the top and bottom 3 categories, Project Goal and Pledge per person. All of the coefficients are significant at 1% level.

	Decision Tree	Logistic Regression
Accuracy	0.718	0.678
Recall	0.745	0.587
Specificity	0.645	0.755
Precision	0.673	0.669

Table-2: Model Evaluation

We compare the Logistic Regression with regression Decision Tree based on measures of classification accuracy. The Decision Tree outperforms Logistic Regression in all measures but Specificity

Conclusion

While there has been work on the Crowdfunding dynamics, the literature is still in its infancy. This project builds on the belief that there is still scope for bringing new perspective in understanding project success in this context. We find that in a logistic regression setting, project category, its goal and the pledge per person have can explain a projects odds of being successful in raising funds through a crowdfunding website. We evaluate the model and compare it with findings from regression Decision Tree. We find that regression decision tree does slightly better in predicting success vs failure of raising funds. This indicates that there might be other factors that can help explain what affects success.

Acknowledgements

We would like to thank our supervisor Olga Baysal for her valuable guidance and motivation which was essential for the progress of our project.

Reference

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. Journal of business venturing, 29(1), 1-16.

Kaggle (2018). Kickstarter projects [Data file]. Retrieved from https://www.kaggle.com/kemical/kickstarter-projects

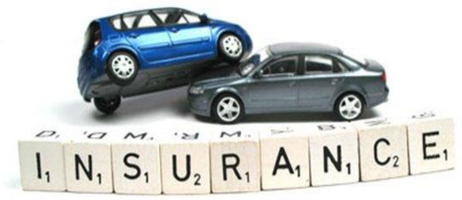
Safe Driver Prediction

Das Moitry , Gourav Mitra

Carleton University

Introduction

This project is about building a model defining the effect of socio-economic factors and technical factors on the accidents that occurred and can predict the probability that the driver will again go through accidents and initiate auto-insurance claim in a given year



Motivation

- Helping Insurance company determine correct insurance plan for drivers.
- Raising safety awareness among drivers.

Data Source

kaggle

40,000 rows of data .

It is a running competition about predicting the probability that a driver will initiate an auto insurance claim in the next year.

Methodology

Data Pre-Processing

- Omitted missing value through R function and deleted irrelevant data.
- Applied algorithms to maintain same Residual Deviance and minimum AIC.

```
***  means 99.9% confident
**   means 99% confident
*    means 95% confident
.    means 90% confident
```

The independent variables not having any stars those were deleted from the model creation function and observed whether Residual deviance remains the same and AIC decrease or not. If so , then those variables were excluded indicating that those variables do not have any effect on our results.

Data Analysis

Classification

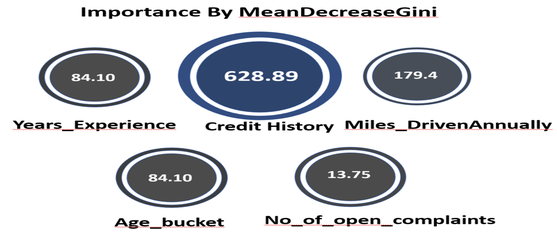
- Decision Tree
- Random Forest
- Logistic Regression

Regression

- Linear Regression
- Poisson Regression

Results

Random Forest model:



That means according to our data , credit history turned out to be the most significant variable in determining whether driver will claim vehicle insurance or not

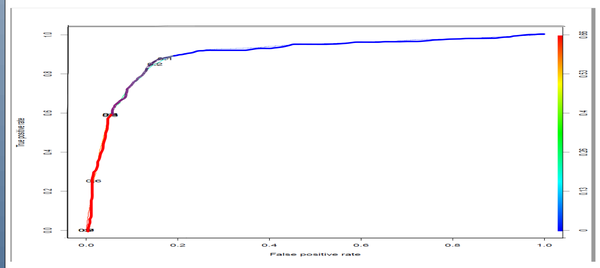
	Predicted Value	
Actual Value	No	Yes
No	2674	181
Yes	96	192

Model predicting the driver will claim auto insurance or not in future. Accuracy: 90%

OOB estimate of error rate: 9.41%.

Therefore, the data which were not in the boot strap and testing , error in indicating the 'yes' or 'no' claim for those data were 9.41%.

Logistic Regression



ROC curve showing threshold 0.1
Accuracy: 0.9008866
Threshold:0.5
Accuracy: 0.710987
Threshold:0.1

	Predicted Value	
Actual Value	No	Yes
No	2300	485
Yes	39	334

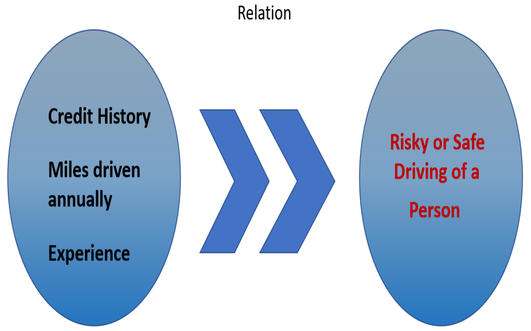
Poisson Regression Model:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.203e+00	5.407e-01	-7.773	7.66e-15 ***
credit_history	-1.553e-04	2.499e-04	-0.621	0.534275
Years_Experience	-6.537e-02	5.076e-03	-12.878	< 2e-16 ***
Miles_driven_annually	-1.469e-06	7.888e-07	-1.862	0.062621 .
credit_history_bucketFair	3.990e+00	5.032e-01	7.929	2.20e-15 ***
credit_history_bucketGood	1.952e+00	5.047e-01	3.869	0.000109 ***
credit_history_bucketVery Good	1.699e+00	5.097e-01	3.333	0.000860 ***
credit_history_bucketVery Poor	4.849e+00	5.089e-01	9.529	< 2e-16 ***

- One unit increase of credit history ,will decrease count of annual_claims by exponential of -1.553e-04 times
- Credit_history_bucket Excellent is kept default which means keeping everything unchanged if credit_history becomes Fair instead of Excellent then claim will increase by exponential of 3.99 times
- P-value : 0.6182274

Conclusion



References

[1] Dan Huangfu . "Data Mining for Car Insurance Claim Prediction" . In Proc. of the 31st IEEE Int. Conference on Software Maintenance and Evolution (ICSME), 2015.

[2] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan, "A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market," In Proc. of the 38th ACM/IEEE Intl. Conference on Software Engineering(ICSE), Austin, TX, May 2016.

[3] S.S.Thakur, J.K. Sing, "Prediction of Online Vehicle Insurance System using Decision Tree Classifier and Bayes Classifier – A Comparative Analysis". In International Journal of Computer Applications (0975 – 8887) International Conference on Microelectronics, Circuits and Systems (MICRO-2014)

[4] S.S.Thakur, J.K. Sing, "Mining Customer's Data for Vehicle Insurance Prediction System using Decision Tree Classifier". In International Journal on Recent Trends in Engineering and Technology, Vol. 9, No. 1, July 2013

Applications of Machine Learning in the Bacterial Genomes of Cystic Fibrosis Patients

Katie Noah[1] and Calvin Jary[2]

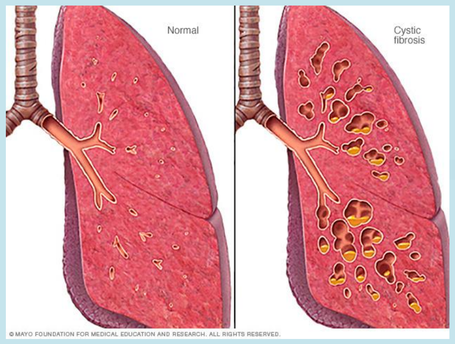
Supervised by: Dr. Alex Wong and Dr. James Green

[1]Department of Biology, Carleton University. katie.noah@carleton.ca [2]Department of Systems and Computer Engineering, Carleton University. calvin.jary@sce.carleton.ca

Introduction

Cystic Fibrosis (CF)

Cystic fibrosis is one of the **leading fatal genetic diseases** [1]. People with cystic fibrosis are immunocompromised and are afflicted by **chronic lung infections**. There is **no cure** for cystic fibrosis, only treatment of symptoms and infections.



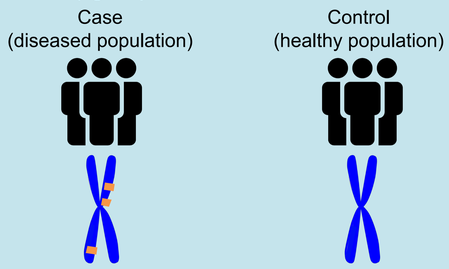
P. aeruginosa

Pseudomonas aeruginosa is an opportunistic bacteria that is the main pathogen in chronic lung infections of CF patients. It undergoes micro-evolutionary changes in a CF lung, such as:

- **Mucoidy** (more mucus like)
- **Biofilm** formation
- Multidrug **antibiotic resistance**
- Adapted **metabolism**
- Higher mutation rates

Gene-association studies

Genome-wide association studies (GWAS) find genes in common between case (diseased) groups, that are not in common with control (non-diseased) groups.



Objectives

- **Identify new genes of interest**
- **Improve and prolong** the lives of people living with *Pseudomonas* infections
- **Provide evidence** for the utilization of machine learning methods in bacterial gene association studies, to obtain **more accurate and complex results**

Methods

Data Collection & Cleanup

- **1586 genomes** from the Pseudomonas Genome Database [2]
- **15,557 genes** total
- **Identify** orthologous genes

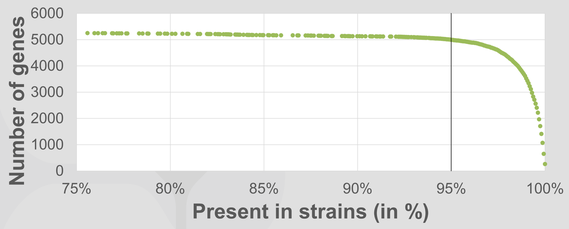


Figure 1. Amount of PAO1 genes present in percentage of strains. 4997 genes appeared in 95% of strains, and was identified as the core genome.

Results

Principal Component Analysis

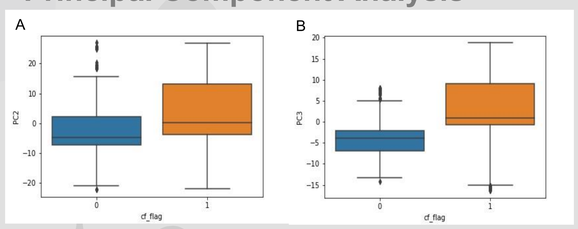


Figure 2. Boxplot of mean principal component values for cystic fibrosis and non cystic fibrosis strains, A) PC2 mean values, B) PC3 mean values.

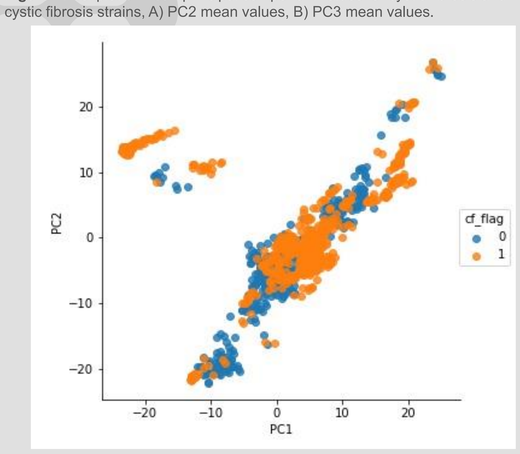


Figure 3. Plot of principal component 2 values against principal component 1 for cystic fibrosis and non-cystic fibrosis *Pseudomonas* strains

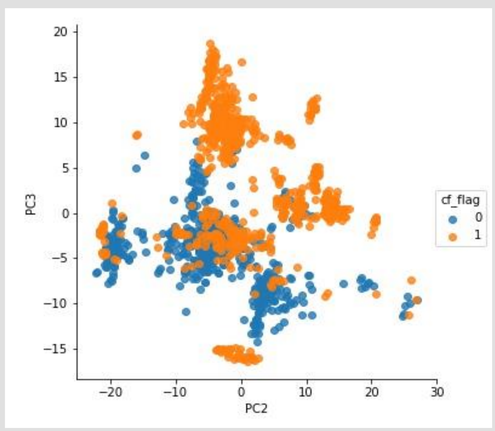


Figure 4. Plot of principal component 3 values against principal component 2 for cystic fibrosis and non-cystic fibrosis *Pseudomonas* strains

Results - continued

Decision Tree

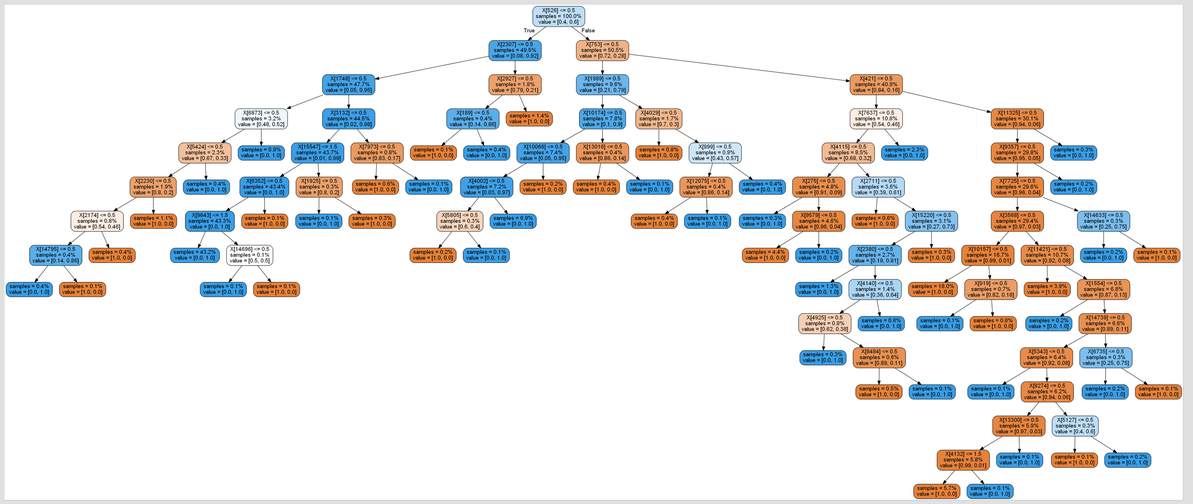


Figure 5. Entropy based decision tree for over 100 genes that separate cystic fibrosis strains from non-cystic fibrosis strains, left to completion.

Fisher's Exact Test

- **>1000 genes** had p-values < 1E-20
- **~200 genes** had odds ratios of infinity

Pearson Correlation

Table 1. Top five gene candidates based on Pearson coefficient

Gene	Pearson coefficient	P-value
hypothetical protein PA0532	0.6502	3.13E-191
hypothetical protein PA0086	0.5812	6.20E-144
flavin-containing monooxygenase	0.5606	6.44E-132
hypothetical protein PA5566	0.5637	1.01E-133
glyceraldehyde-3-phosphate dehydrogenase	0.5590	5.04E-131

Random Forest

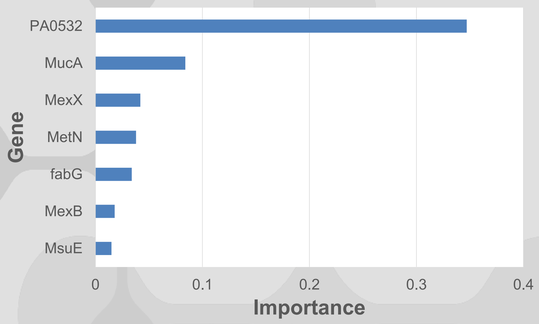


Figure 6. Top genes that contribute to the CF strain genotype, using random forest

Support Vector Machines

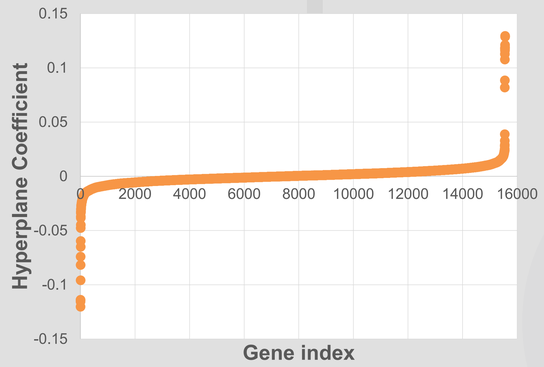


Figure 7. Top genes that contribute to the CF strain genotype, using random forest

Conclusions

Biological

- **Hypothetical protein PAO0532**
- **Mucoidy** (*mucA*)
- **Antibiotic-resistance** (*mexX, mexB)*
- **Metabolism** (*metN, fabG* and *msuE*)

Methodological

- **Random forest** gave the most useful results
- Found **231 genes** that were important in characterizing CF strains

Future Work

- **Mutations within core genes** are being characterized for analysis.
- Characterization of important hypothetical protein coding genes

Limitations

- Not corrected for phylogenetic/population structure

Acknowledgements

Thank you to the Wong and Green labs for the advice and guidance along the way. Inspiration for the project came from Dr. Thienny Mah, at the University of Ottawa.

References

[1] Cystic Fibrosis Canada. (n.d.). Retrieved from http://www.cysticfibrosis.ca/about-cf

[2] Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., & Brinkman, F. S. (2016). Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. Nucleic Acids Research.

Carleton UNIVERSITY

Large-Scale Land Cover Classification with Convolutional Neural Networks

Ekaba Bisong and Yajing Deng
School of Computer Science & Sprott School of Business,
Carleton University
Supervisor: Elio Velazquez

Abstract

This study designs a large-scale convolutional neural network to build a land cover classifier using the SAT-6 airborne dataset made available by the National Agriculture Imagery Program (NAIP). We discuss the impacts of incorporating state-of-the-art vision models into existing Remote Sensing and Geographic Information Systems pipelines with particular emphasis in its role as drivers of environmental and economic growth in rural areas/ developing communities. This project leverages the Google Cloud Platform infrastructure.

Introduction

Land cover classification is the process of identifying land cover types from remotely sensed satellite imageries by operating on pixel information in digital maps.

Figure 1: Land cover classification comprises of supervised, unsupervised and object-based techniques for identifying land cover types. The supervised method of land classification is the focus of this work - due to its more robust evaluation mechanisms for determining the performance of the learning model.

Methodology

Convolutional Neural Networks (CNN) have emerged as the state-of-the-art technique for automatic object classification and image recognition tasks by stacking deep neural semantic layers to capture sophisticated feature representations from image data.

Model Architecture

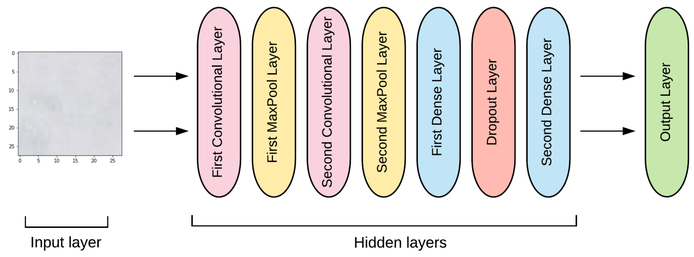


Figure 2: The CNN architecture contains 2 convolutional layers, 2 pooling layers, 2 dense layers, with one dropout layer in-between for regularization.

Model Hyper-parameters

- Conv. Layer #1: A 32, 5x5 filter (extracting 5x5-pixel subregions), with ReLU activation function.
- Pooling Layer #1: Max pooling with a 2x2 filter and stride of 2 (which specifies that pooled regions do not overlap).
- Conv. Layer #2: Applies 64 5x5 filters, with ReLU activation function.
- Pooling Layer #2: Another max pooling with a 2x2 filter and stride of 2.
- Dense Layer #1: 1,024 neurons, with dropout regularization rate of 0.5 (probability of 0.5 that any given element will be dropped during training)
- Dense Layer #2 (Logits Layer): 6 neurons, one for each land use class.

Training Infrastructure

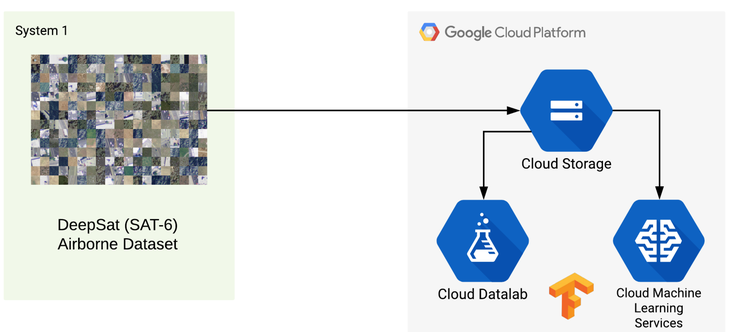


Figure 3: Designed model on Datalab & distributed training with Cloud ML

Results

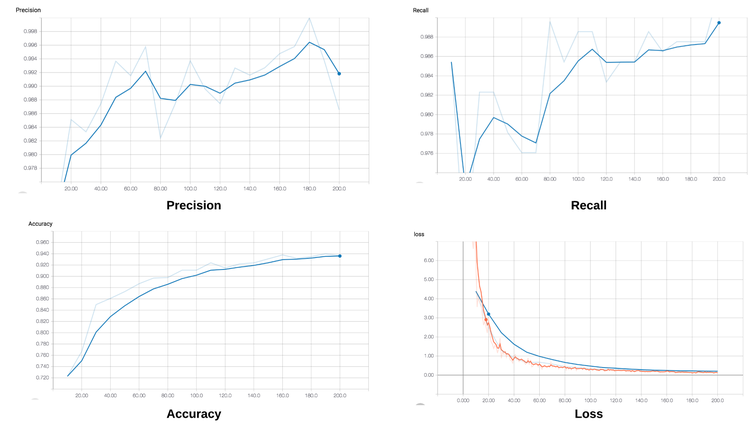


Figure 4: An increasing precision and recall indicates the prediction is both relevant and correct. Also, observe the increase in accuracy and the decreasing loss on both the training (orange) and evaluation (blue) datasets.

Significance/ Implication

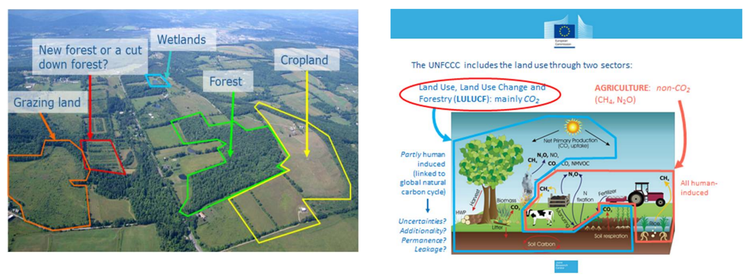


Figure 5: Remote sensing is a valuable tool for identifying land use and land cover changes. The analysis of remotely sensed data provides critical insights into the evolving human-environment relationship. In particular, the analysis of multispectral imagery is a key driver for estimating greenhouse emissions from Land Use, Land Use Change and Forestry (LU-LUCF)

References

[1] Saikat et. al. Basu. Deepsat: a learning framework for satellite imagery. In *Proc. of the 23rd SIGSPATIAL Intl. Conf. on Adv. in GIS*, page 37. ACM, 2015.

Identifying poisonous mushrooms based on morphological characteristics

W.M. Twardek[1], & A. Hill[2]

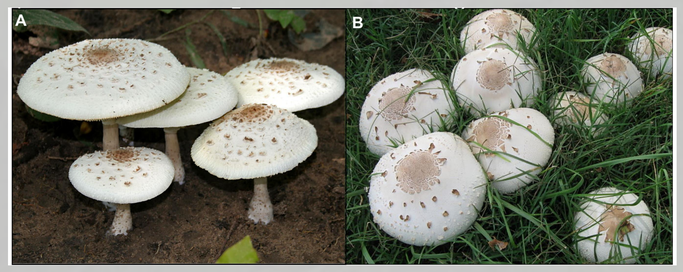
Carleton UNIVERSITY
Canada's Capital University

[1] Department of Biology, Carleton University, Ottawa, Canada
[2] Department of Engineering, Carleton University, Ottawa, Canada

Data Day 5.0
March 27, 2018

INTRODUCTION

- Mushroom picking is a popular recreational activity across Europe, parts of Asia, and North America [1,2].
- In Czech Republic, mushroom picking occurs in 75% of households and is valued at over 180 million dollars annually [3]
- Consumption of wild mushrooms represents an unregulated food source that could pose human health risks.
- Mushrooms maintain diverse chemical properties that can be poisonous to humans, resulting in serious health conditions including liver and kidney failure among others [4]. [8]
- For example, just 25-30 grams of the Death Cap mushroom (Amanita phalloides) can be fatal [5].
- In France alone it is estimated that 8,000-10,000 people suffer from mushroom poisoning each year [6].
- The rich diversity of mushroom species (approximately 14000 species described) makes it difficult for an individual mushroom picker to identify all potentially inedible species [7].
- Textbooks and classification keys can only go so far, people need an automated and reliable way to inform their choices.



METHODS

- OBJECTIVES:

1) To predict mushroom edibility based on mushroom morphological characteristics
2) To determine the fewest number of features needed to minimize classification error to <1%

- Random decision forests were used to classify a mushroom as edible or poisonous based on morphological traits
- Manual stepwise inclusion of variables based on their relative influence on the model
- Cost matrix specified greater costs for classifying false negatives (labelling poisonous mushrooms as edible
- The model was trained with a subset of data and cross validation was used to determine classification reliability before testing on the full dataset.
- Model performance was evaluated based off the confusion matrix and a comparison to expected values based off a null model

DATASETS & LIMITATIONS

- Simulated mushroom data from the Audubon Society Field Guide to North American Mushrooms
- 22 categorical morphological traits related to colour, shape, and size of various mushroom structures (~8000 observations)

RESULTS

- 5 variables were needed to design a condensed model with high sensitivity and specificity
- Influence on the model was greatest for odor (0.63), followed by spore print colour (0.20), gill size (0.07), gill colour (0.069), and ring type (0.052)

Table 1. Sensitivity and specificity measures for the full and condensed models predicting mushroom edibleness

Performance	Full model (22 variables)	Condensed model (5 variables)
Sensitivity	0.99	1.0
Specificity	1.0	1.0

- Both the full model (χ^2 =1620, P≃0.0) and condensed model (χ2 =1584.4, P≃0.0) had predicted values that agreed strongly with labelled classes
- There was no significant difference between the predicted values generated from the full candidate model and the condensed model (χ2 =1588.2, P≃0.0)

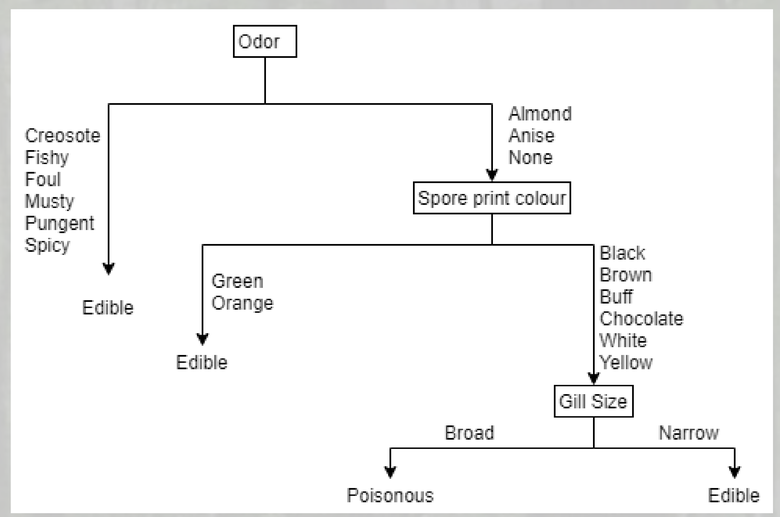


Figure 1. A sample decision tree used to classify mushrooms as poisonous vs. edible. This diagram uses hypothetical data and should not be used to inform mushroom eating decisions.

CONCLUSIONS

- Random decision forests with a specified cost matrix was an effective means of classifying mushrooms
- Findings from this work could be used to develop a classification key, or smartphone application to help mushroom pickers make informed decisions when consuming wild mushrooms
- Future work could evaluate the use of Artificial Neural Networks to predict mushroom edibleness based off user submitted images

ACKNOWLEDGEMENTS

I would like to thank Dr. Elio Velazquez and Dr. Olga Baysal for their feedback on the project.

LITERATURE CITED

[1] Stasys Mizaras, Marius Kavaliauskas, Gintautas Činga, Diana Mizaraitė, and Olgirda Belova. 2015. Socio-economic aspects of recreational use of forests in Lithuania. Balt. For. 21, 2 (2015), 308–314.

[2] C. J. E. Schulp, W. Thuiller, and P. H. Verburg. 2014. Wild food in Europe: A synthesis of knowledge and data of terrestrial wild food as an ecosystem service. Ecol. Econ. 105, (2014), 292–305.

[3] Ludek Sisak, Marcel Riedl, and Roman Dudik. 2016. Non-market non-timber forest products in the Czech Republic-Their socio-economic effects and trends in forest land use. Land use policy 50, (2016), 390–398.

[4] David Varvenne, Karine Retornaz, Prune Metge, Luc De Haro, and Philippe Minodier. 2015. Amatoxin-containing mushroom (Lepiota brunneoincarnata) familial poisoning. Pediatr. Emerg. Care 31, 4 (2015), 277–278.

[5] Christine Karlson-Stiber and Hans Persson. 2003. Cytotoxic fungi - An overview. Toxicon, 42, (2003) 339–349.

[6] P. Saviuc, and F. Flesh. 2003. Intoxication par les champignons et leur traitement. Presse Med. 32, (2003). 1427–1435.

[7] Philip G. Miles, Shu-Ting Chang. 2004. Mushrooms: Cultivation, Nutritional Value, Medicinal Effect, and Environmental Impact. Boca Raton, Florida: CRC Press (2004). 480 pp.

[8] Thomas N. Sherratt, David M. Wilkinson, and Roderick S. Bain. 2005. Explaining Dioscorides' "Double Difference": Why Are Some Mushrooms Poisonous, and Do They Signal Their Unprofitability? Am. Nat. 166, 6 (2005), 767–775.

Building a RNN-Based Prediction Model for Restaurant Visitor Volume

Kevin Hua
Master of Computer Science
Dr. Yuhong Gao

Quan Gao
Master of Business Administration
Dr. Elio Velazquez

Carleton UNIVERSITY
DATA DAY 5.0

Introduction

Problem: Restaurants are facing challenges of under or over buying and staffing, which cause 4%-10% of food inventory throwing away, costs of idle servers and customer dissatisfaction.

Proposition: Provide robust and accurate predictive model for customer volume forecasting via machine learning.

Dataset

Our dataset comes from three separate sources: Hot Pepper Gourmet, AirREGI, and weather stations across Japan. Across a total of 8 CSV files, we pulled in the following data (4700 restaurants and 1600 weather stations):

Store Information
- Store ID
- Genre
- Visitor Count (per day)
- (Latitude, Longitude)
- Region

Date Information
- Day of Week
- Holiday (boolean)
- Calendar Day

Weather Information
- Hours of Sunlight
- Total Precipitation
- Total Snowfall
- Average Temperature
- Average Windspeed

Methodology

Feature Matrix: For each store, generate the following feature matrix, where each row represents a given date.

$$\begin{bmatrix} region_i & weekday_i & genre_i & holiday_i & sunlight_i & precipitation_i & snowfall_i & temperature_i & windspeed_i \\ region_{i+1} & weekday_{i+1} & genre_{i+1} & holiday_{i+1} & sunlight_{i+1} & precipitation_{i+1} & snowfall_{i+1} & temperature_{i+1} & windspeed_{i+1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ region_n & weekday_n & genre_n & holiday_n & sunlight_n & precipitation_n & snowfall_n & temperature_n & windspeed_n \end{bmatrix}$$

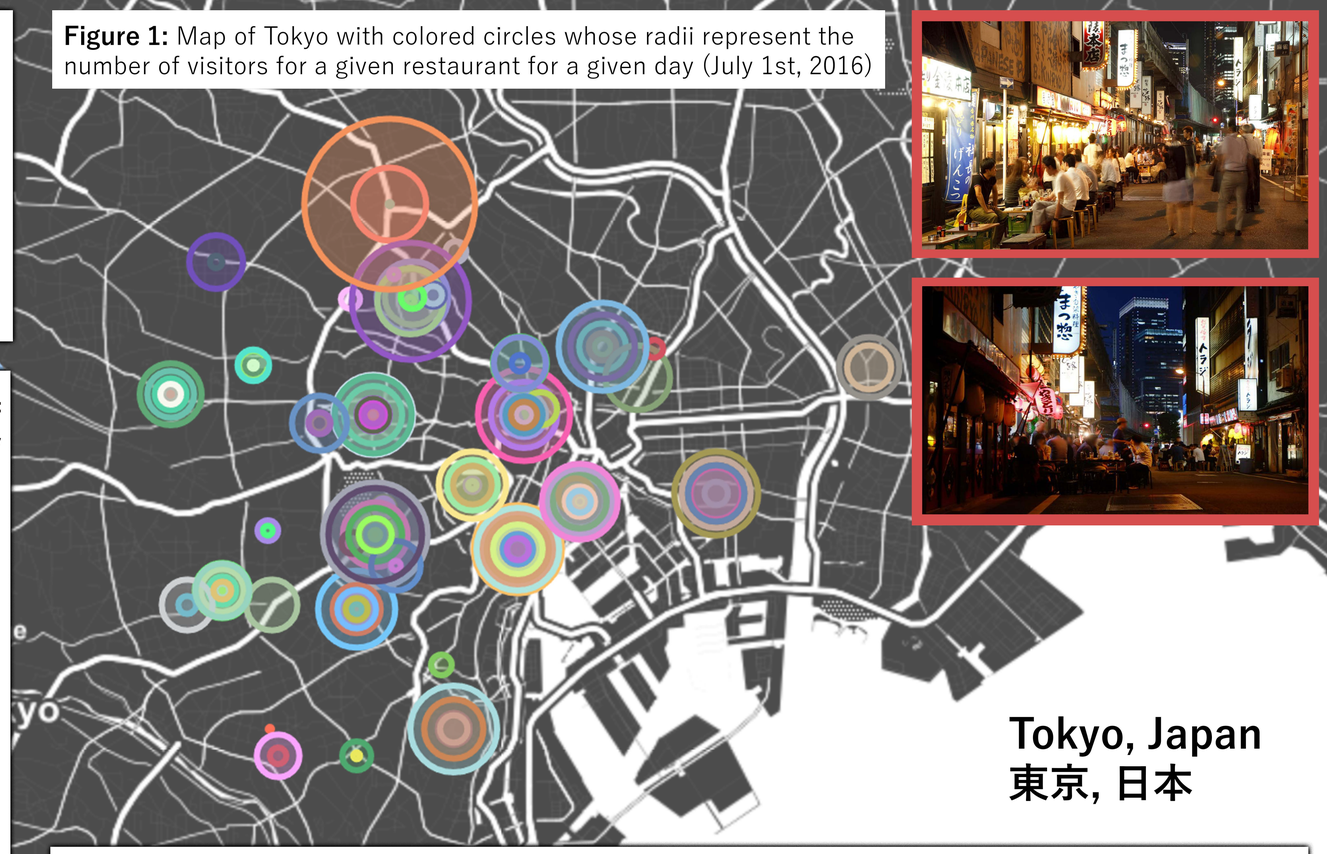
Data Preprocessing Stage Workflow

Extract Data from CSV → Consolidate Data Clean Data → Generate Feature Matrix → Encode Data (Normalize)

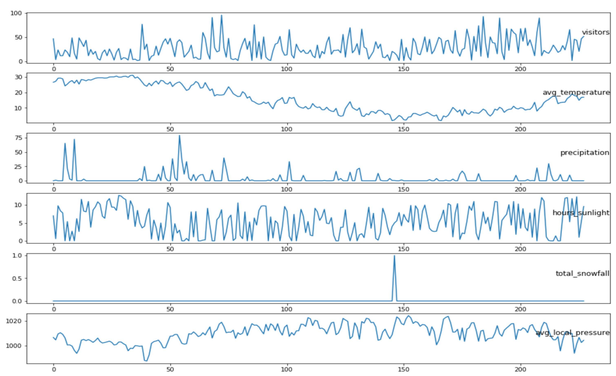
RNN-LSTM Stage Workflow

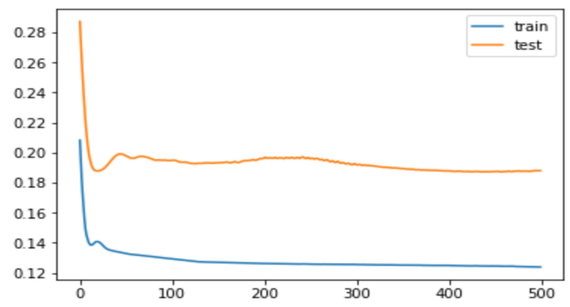
Train Model (1:4) → Generate LSTM Layers → Set Loss Set Optimizer → Test Model

Figure 1: Map of Tokyo with colored circles whose radii represent the number of visitors for a given restaurant for a given day (July 1st, 2016)



Results and Conclusions





Mean RMSE: 0.257282
Mean MSE: 0.068551

Results are promising, but not good enough yet for real-world usage. Stores were measured separately; a method to learn from each other might improve results. More factors to consider that were not included in our datasets include major/minor geopolitical events, construction, competition, ratings, or community events.

References

[1] [n.d]. Calendar for Year 2017 (Japan). ([n.d]). https://www.timetable.com/calendar/?year=2017&country=26
[2] 2018. Recruit Restaurant Visitor Forecast. (2018). https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data
[3] 2018. Weather Data for Recruit Restaurant Competition. (2018). https://www.kaggle.com/huntermcgushion/rrv-weather-data
[4] Shannon Arnold. 2014. Keys to Making Accurate Sales Forecast. (2014). https://www.foodnewsfeed.com/fsr/vendor-bylines/keys-making-accurate-sales-forecasts
[5] Lilian Weng. 2017. Predict Stock Prices Using RNN: Part 1. (2017). https://lilianweng.github.io/lil-log/2017/07/08/predict-stock-prices-using-RNN-part-1.html

Acknowledgements: We would like to thank Kaggle, Recruit Lifestyle, Hot Pepper Gourmet, and AirREGI for the datasets we used.

Household Power Consumption Analysis and Forecasting

Arslan Ahmed[1], Anupam Sehgal[2], Mohamed Ibnkahla[1], Elio Velazquez[3], Olga Baysal[3]
[1]Dept. of Systems and Computer Engineering , [2]Sprott School of Business, [3]School of Computer Science
Carleton University

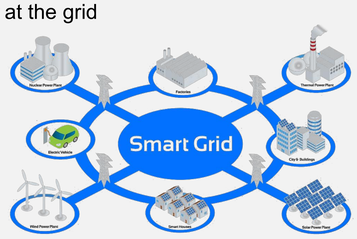
Introduction

What is Demand Response in Smart Grid?

Negotiation with the consumers to shift their electricity usage during peak hours in response to time-based rates or other incentives

Why is it important?

- Balance electric demand and supply at the grid
- Reduce consumer's monthly bill
- Reduce carbon emissions

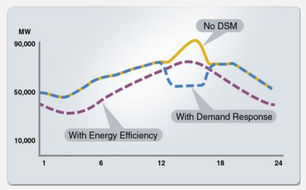


Challenges:

- Accurate prediction of peak hours
- Consumer profiling
- Awareness and understanding of the consumer about his load pattern

Objectives

- Explore the effects of the following on household electricity consumption:
  - Consumer's demographics
  - Static characteristics
  - Weather
  - Dynamic time-of-use (dToU) pricing
- Forecast short term energy consumption
- Make recommendations to the utility and consumer



Dataset

Project:
Low Carbon London Trial (LCL)

Location: London, UK

Time period: 2013

Households: 5567

dToU pricing: 1122

Smart Energy Meter Readings
(half-hourly)

Static characteristics
- Type of house
- Insulation material
- No. of rooms/occupants
- Type and number of appliances etc.

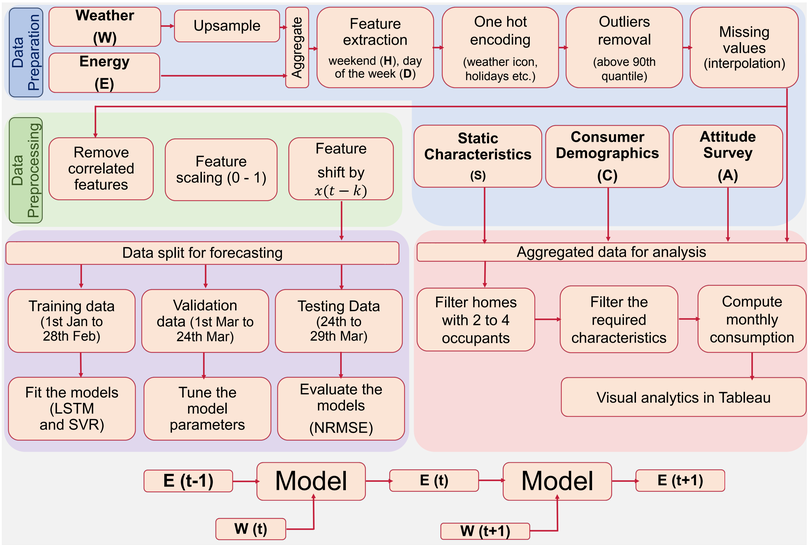
Consumer demographics
- Age category
- Income level
- Gender etc.

Survey data
- Attitude of consumers towards dToU

Hourly weather data
- Source: Dark Sky API
- Temperature
- Pressure
- Humidity
- Visibility
- Wind-speed

Limitations:
- Bulk of Missing categorical data for static characteristics and demographics
- Frequently changing weather data of London

Methodology



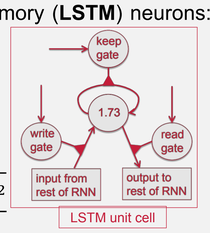
Techniques

- Recurrent Neural Networks (**RNN**) with Long-Short-Term-Memory (**LSTM**) neurons: 2 hidden layers, Adam Optimizer
- Support Vector Regression (**SVR**) with Gaussian Kernel

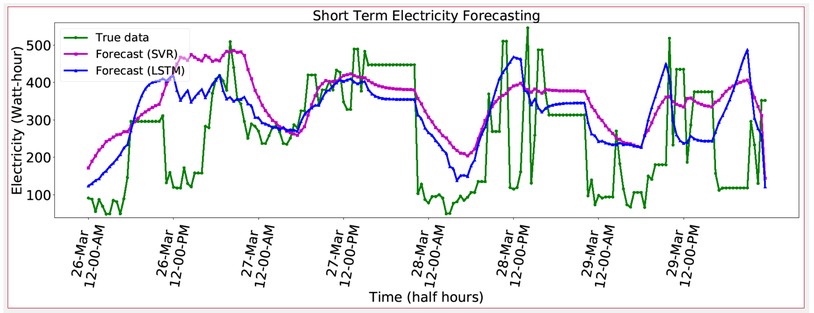
$$K(x,y) = exp(-\frac{||x-y||^2}{2\sigma^2})$$

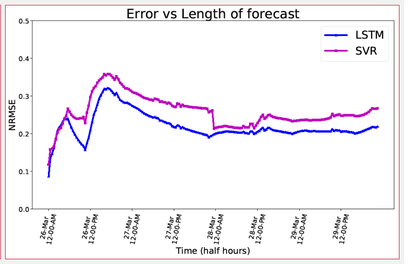
- **Evaluation metric**: Normalized Root Mean Square Error

$$NRMSE = \frac{1}{y_{max} - y_{min}} \times \sqrt{\frac{\sum_{t=1}^{n}(y_p - y)^2}{n}}$$

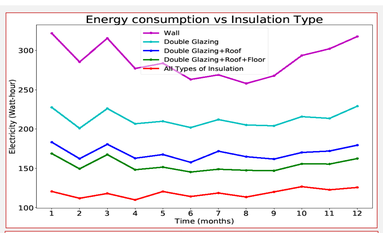


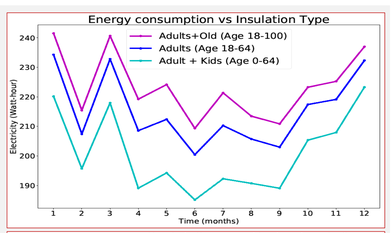
Results

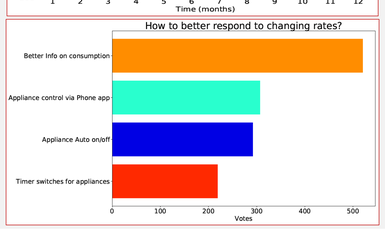


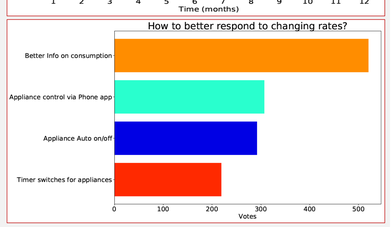


$NRMSE_{LSTM}$: 0.21 $NRMSE_{SVM}$: 0.28









Conclusions

- For short term forecasting:
  - LSTM networks perform slightly better than SVR
  - Small amount of training data is well sufficient
  - Error increases upon increasing the forecasting period
  - Weather parameters are not highly weighted
- Consumers can save up to 30.5% energy by using full insulations instead of the commonly used double glazing insulation
- Utilities can get better result out of dToU program by keeping the consumers well-informed about their consumption, and making the dToU more predictable e.g. every Sunday

Future Work

- Discover unique patterns in each house for real time personalized dynamic energy pricing

References

[1] A. Ahmed, K. Arab, Z. Bouida, and M. Ibnkahla, "Data Communication and Analytics for Smart Grid Systems," *Proc 2018 IEEE International Conf. on Communications (ICC'18)*, Kansas City, MO, USA, pp. 1-6, May 2018.

[2] Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid*.

[3] Rodrigues, F., Cardeira, C., & Calado, J. M. F. (2014). The daily and hourly energy consumption and load forecasting using artificial neural network method: a case study using a set of 93 households in Portugal. *Energy Procedia*, 62, 220-229.

[4] Hinton, Geoffrey. "Neural Networks for Machine Learning, Coursera" Retrieved March 27, 2018, from www.coursera.org/learn/neural-networks

[5] U.S Dept. of Energy. "What is the Smart Grid?" Retrieved March 27, 2018, from https://www.smartgrid.gov/the_smart_grid/smart_grid.html