Making Exploratory, Text and Predictive Analytics Accessible to Everyone

Shiny PET: A Predictive, Exploratory and Text R Shiny Application using Airbnb Data

Ang Su Yiin ¹ Suyiin.ang.2020@mitb.smu.edu.sg Joey Chua ¹ joey.chua.2020@mitb.smu.edu.sg Kevin Gunawan Albindo ¹ Kaalbindo.2019@mitb.smu.edu.sg Kam Tin Seong ¹ Singapore Management University

Introduction

With the explosion of affordable data storage and processing technologies, the **demand for data-driven decision-making** (DDDM) has increased significantly. Although there is an extensive range of commercial statistical tools, they are often **subscription-based** and **demand good technical knowledge** to mine and draw insights from. Therefore, it may not appeal to the average user.

Motivation

The motivation of this project stems from two main issues - the **proliferation of data** and **lack of user-friendly open source tools** to make data-driven decision.

Exploratory Geospatial Analysis

- Choropleth map - shows summary of listings by Singapore neighbourhood.

- Point symbol map - shows the distribution of listings throughout Singapore.



Exploratory and confirmatory analysis

- Panel on the left allows user to seamlessly toggle between different charts.

- Automated statistical results based on selected variables.

- Interactive plots - user can select, pan, zoom and download plot.



text,we can observe the connections between words. To make visualisation interpretable, only the top 100 words are shown.
Correlation Network - Users can observe words that tend to co-occur within particular documents/topics even if they don't exist in the same sentence.



Predictive

Data splitting

- Proportion of training data are customisable with distribution plot between train-test set displayed for awareness of potential bias.



Feature selection

- Interactive correlation matrix plot with options for different p-

The aim is to create an R Shiny application to **encourage anyone and everyone to make data-based decisions effortlessly**.

Approach



Application modules and flowchart





Text

Token Frequency

- Word Cloud allows visualisation of qualitative data by showing the frequency of appearance of a particular word in Airbnb reviews.

Bar chart sorts the frequency of word in descending order.Unigram/Bigram allow user to understand reviews in various

forms.

requency Sentiment analysis Topic Modelling Network analysis	
ord Cloud	Word Frequency
Uni-gram 🔍 Bi-gram	
0 0	stay
chinatown neighborhood	clean
-nient Comfortable	location
converie cindanoro -	nice
Lathroom SIIICADUIC esta	host
breakfast highly excellent	mrt
	singapore
	recommend
stance airbnb	apartment
	comfortable
	close
clean 12817 Superday bed null easy night	easy
hort 9 guiet	helpful
	minute
	time
staff space lots pool minute	friendly
hostel	check
perfect	station
amazing US sport and a comperience	walk
Contraction Pleasant facilities	food
	convenient
wonderful hospitality responsive	0 2.5k 5k 7.5k 10k 12.5k 15k 17.5k 20k

Sentiment Analysis

- Word Clouds are based on sentiments leixcons.

- For AFINN, the mean score and frequency of words are shown in either positive (>0) or negative (<0).

- For BING, the count of the words are shown.

- FOR NRC, radial plot to show the tendency of sentiments is shown.



Topic Modelling

- Topic Modeling help users interpret the topics that has been fit to a corpus of Airbnb data.

- The left panel shows a global view of topic model to show the prevalence of each topic, and how topics relate to each other.

- The right panel shows horizontal barchar where the length of the bars represent individual terms that are most useful in interpreting the selected topic. This enable users to understand the meaning behind each topic. value and correlation method.

- Comparison of variable importance from Random Forest and Boruta method.



Variable selection and transformation

- Multi-input form for variable selection and transformation steps information display.

- Pre and post transformation data distribution plot for numerical variables.



Model training and validation

- A selection of model types to choose for developing predictive model with interactive plots for training and validation results.

- Prediction error evaluation by comparing training data distribution and high error cases, with top predictors selected by user.



Model selection

- Comparison of models based on performance metrics with option to select the best model for prediction.

- Using the selected model, predictors value can be adjusted to predict the target variable.



Exploratory

Summary of dataset

- Interactive table allows selection and filter of variables

Observe variables 1 VARIABLES 48	les		i≡	observations 3970			Ģ	FACTOR VARI	ABLES			NUMERICAL VARIAN	SLES
Summary of factor v	aria	bles											Search:
		skim_variable	¢.	n_missing (complete_rate (ordered		0		n_unique	top_counts	\$
		All	All		All		All			All		All	
		1 host_response_time		0		1	false					5 wit: 1812, wit: 846, N/A: 491, a f: 4	176
		2 host_is_superhost		0		1	false					2 FAL: 3297, TRU: 673	
		3 host_identity_verified		0		1	false					2 TRU: 3304, FAL: 666	
		4 neighbourhood_cleansed		0		1	false				4	2 Kal: 557, Gey: 374, Out: 367, Nov: 356	
		5 neighbourhood_group_cleansed		0		1	false					Cen: 3238, Eas: 257, Wes: 230, No 135	16.
Showing 1 to 5 of 9 entries													Previous 1 2 Next
Summary of numeric	cal	variables									Search:		
		skim_variable	n_missing (complete_rate (mean (si	1¢	po 🗧	P	50 🛊	p100 0		
		All	All	All	1	All	All		All	All			
	1	review_scores_location	1579	0.6	9-44	0.9	94	2		10	10		
	2	review_scores_value	1579	0.6	9.04	1.	17	2		9	10		
	3	review_scores_checkin	1578	0.6	9-59	0.9	95	2		10	10		
	4	review_scores_accuracy	1577	0.6	9.3	1.	15	2		10	10		
	5	review_scores_communication	1577	0.6	9.57	0.9	8	2		10	10		
Showing 1 to 5 of 39 entries									Previou	5 1 2	3 4 5	8 Next	

Topic Modelling



Network Analysis

- Bidirectional network - Utilising the concept of nodes, edges and

Further work

- Include a data load and wrangling function to accommodate different datasets.
- Combine text and explore modules with coordinated linked views for multi-dimensional exploration.
- More predictive models such as neural network and classification workflow can be included to provide user with wider predictive analytics selection.



