Capstone Project:
London Transport Policy Analysis

Part 1. Overview

Executive Summary

This project investigates whether certain transportation policies can have a discernible impact on land use values and firm locations over time in a city. This was the topic of my doctoral dissertation at UC Berkeley in 2015. Using average travel time by car and public transit as transportation accessibility indicators and urban economic data such as residential and firm populations and rents, binary classification models are estimated to understand the features most associated with areas located inside the congestion charge zone, compared with areas outside, and the most significant drivers of change over time.

1. Research Understanding

Rationale

There is a great deal of political resistance to the idea of reducing vehicle traffic in cities via congestion charging policies. If the policy is merely a tax on drivers which lacks benefits to the urban economy, this is correct. However if there are are demonstrable economic benefits in terms of employment growth and transit quality improvements, this would be an important finding in the field of urban planning.

1.1 Background

Congestion charging defines a zone or set of road corridors where drivers are charged usage fees during peak congestion hours. A congestion charging zone was implemented in London in 2003. It is a ring around the historic downtown area where the bulk of firms, governmental agencies, and cultural amenities are located. Vehicles that drive across the perimeter were initially charged a fee of £5 to enter the downtown zone. The congestion charge fee was increased to £8 in 2008 and £10 in 2011. It congestion charge was considered an instant success when implemented: 20% fewer cars drove into central London, meaning reduced traffic congestion and faster travel times by bus.

The congestion charge was paired with extensive improvements to the public transport system. For example, many new buses were added to the fleet, and service frequencies were increased for both bus and rail services. Many studies have established the impact of public transport accessibliity on land values, over time. For example, rents typically increase rapidly within a half-mile area around new commuter rail stations after they are built. In the lingo of urban economics, the accessiblity benefits of public transport are capitalized into higher land values. A secondary impact of higher rents resulting from accessibility benefits is firm relocation. Only the firms that highly value accessiblity are willing to pay higher rents for it. For example, large firms with many employees, or firms whose customers are other firms that value locating in urban agglomerations. Since congestion charging is an accessibility benefit similar to public transport improvements, should it have a similar observable effect? The research goal is to understand whether the congestion charging zone has discernible differences in urban economic data, after nearly a decade in operation.

The project uses uses a dataset that I created from data sources in London, England during fieldwork in 2013 and 2014. Economic data was collected for 4765 small geograpic areas in London called LSOAs, which are a similar size as electoral wards. Data for each LSOA was collected for the year 2001 and 2011, including counts of people and firms, average travel times by car and public transport, and average rent levels for office and retail space.

1.2 Research Objectives

This project seeks evidence of changes in the urban economy of downtown London during a ten-year period after public transport improvements were introduced along with a congestion charging zone. If no evidence is found, the null hypothesis that these transport policies have no discernible impacts on land use will be supported.

Hypothesis

Public transit improvements paired with congestion charging produced accessibility benefits within the congestion charging zone which were capitalized into higher residential and commercial land rents. There were two primary drivers of change, lower travel times by public transit and increasing commercial rents. Improved accessibility by public transit made the zone more attractive to employers, meaning the population of firms grew and jobs became more concentrated there. Certain industries that were growing during the timeframe of the study became more concentrated in the zone, particularly creative and professional industries whose productivity relies upon employee interactions and knowledge spillovers, and business services industries whose customers are other firms. Other industries that were shrinking or did not receive productivity benefits were less able to keep up with rising commercial rents and became less concentrated in the zone, such as

wholesalers and manufacturers. Higher commercial rents made the zone less attractive to others, in particular employers with vehicle fleets, small employers, and shrinking industries became less concentrated. Meanwhile higher residential rents kept the residential population stable or shrinking.

Research Questions

In the congestion charge zone, compared to other areas of London:

- Has the population of people grown or shrunk at a faster or slower rate?
- Has employment in the congestion charge area grown or shrunk at a faster or slower rate?
- Have average access times by car and public transit become faster or slower?
- Have office, retail, and warehouse rents increased or decreased?
- Have the number of large, medium, small and micro firms increased or decreased?
- Which industries became more concentrated inside the zone, and which less concentrated?

Constraints

The two transport policies, public transit improvement and congestion charging, were paired together, making it difficult to disentangle the effects of each measure. Since we do not have a counterfactual case, or London without a congestion charge zone, it is impossible to isolate the impacts of these transport policies from other drivers of change and attribute the findings exclusively to them.

1.3 Methodology

This project applied classification and supervised learning methods to address the research questions. Models were estimated for three datasets representing 2001, 2011 and percent change between those years. Results were compared between datasets. An improvement in classifier predictive power over time, as evidenced by higher metrics, was interpreted as evidence that the congestion charge zone has become more differentiated from other LSOAs over time.

Exploratory Analysis:

- Feature distributions and correlations
- Correlations with target feature, congestion charge zone
- Cluster analysis
- Principal component analysis

Modeling:

- A logistic regression model identifying the most important features associated with LSOAs inside and outside the congestion charging zone in 2001, in 2011, and with percent change over time
- A Random Forest binary classification model capable of sorting LSOAs into two classes, inside and outside the congestion zone, which has stronger predictive power in 2011 than 2001
- A Neural Network binary classification model capable of sorting LSOAs into two classes, inside and outside the congestion zone, which has stronger predictive power in 2011 than 2001

Exploratory Data Analysis

2. Data Description

Since this project seeks evidence of change over time, it uses a panel dataset with features representing the cases (LSOAs) at two points in time, and change over time:

- Before the congestion charge was implemented, in 2001
- After the congestion charge was implemented, in 2011
- Percent change over this period

The target variable "cc" is a binary class representing whether the geographic unit (LSOA) is located inside the congestion charge zone or not.

The unit of analysis is the LSOA, or lower super output area, a statistical geography used by the city of London. The dataset contains 4765 of these, along with about 25 features for each. Each LSOA is created to contain 400-1200 households or 1000-3000 people. Thus they vary in geographical size but are normalized for population. For this reason, this project does not normalize the data for spatial area, that is, it uses population counts per LSOA rather than population density. More details about LSOAs can be found on the UK Office for National Statistics website, https://www.ons.gov.uk/methodology/geography/ukgeographies/statisticalgeographies

2.1 Data Sources

The sources for the features described in the data dictionary below are as follows:

- Population data, UK Census

- Employment data and travel times, UK Department for Transport Accessibility Statistics
- Firms data, UK Office for National Statistics, Business Structure Database (microdata)
- Rents data, UK Valuation Office Agency, Commercial Floorspace Rateable Value Statistics

It is important to note that the data used to represent commercial rent levels is not sourced from actual commercial rents, but from tax ratings data. This proxy was used because it is publicly available and consistently calculated land valuation data.

I created this dataset from public data sources in London, England during doctoral fieldwork in 2013 and 2014. Economic data was collected for 4765 small geographic areas in London called LSOAs, which are a similar size to electoral wards. Similar data for each LSOA was collected for two years, 2001 and 2011, including counts of people and firms, average travel times by car and public transport, and average rent levels for office and retail space. The methodology is fully described in the UC Berkeley dissertation, https://escholarship.org/uc/item/0px3f6gk

2.2 Data Dictionary

The initial dataset had a total of 72 features.

Feature name	Units	Definition
lsoa01		LSOA is a UK geographical unit "lower super output area" representing 400-1200 households or 1,000 to 3,000 people, defined in 2001
lsoa01_name		LSOA name
lsoa_area	Hectares	LSOA area
cc		Indicator is 1 if LSOA is located inside the London Congestion Charge zone and 0 if not
pop_2001	People	Population count
pop_2011	People	Population count
pop_pct_chg	Percent	Percent change in population counts from 2001 to 2011
Note – each feature below has 2001, 2011, and percent change values, named in the same convention as the population feature above		
jobs	Jobs	Employment count
car_time	Minutes	Average minimum travel time by car to this LSOA
pt_time	Minutes	Average minimum travel time by transit to this LSOA
rent_off	Value per sq meter	Office space rent proxy, the tax rateable value in 2001
rent_ret	Value per sq meter	Retail space rent proxy, the tax rateable value in 2001
rent_whs	Value per sq meter	Warehouse space rent proxy, the tax rateable value in 2001

large_firms	Firms	Count of firms with 501 to 1000 employees
med_firms	Firms	Count of firms with 101 to 500 employees
small_firms	Firms	Count of firms with 11 to 100 employees
micro_firms	Firms	Count of firms with 1 to 10 employees
afm	Firms	Agriculture, fishing, mining, other natural resources
bizsup	Firms	Business support services, legal, accounting,
comtelrd	Firms	Computer hardware & software, telecom, research & development
creative	Firms	Publishing, advertising, media, architecture
cult	Firms	Culture, libraries, museums, sports, hotels, restaurants, theatre
devel	Firms	Real estate & construction, hardware, building supply
eduhsw	Firms	Education, health and social work
mgmt_	Firms	Management consulting
pubutil	Firms	Public utilities and administration
retail	Firms	Retail trade
tsp	Firms	Transportation
ws	Firms	Wholesale trade

2.3 Data Cleaning

Data cleaning and exploratory data analysis was performed in a Jupyter notebook which can be found here, link:

- All the features are numerical so there is no prep work to be done with strings or categoricals
- No duplicates were found
- Many features have highly skewed distributions, therefore all data was normalized to be on the same scale
- Several features had outliers, such as a high concenrtation of Jobs in downtown LSOAs, but since they were accurate data they were left in the dataset
- Many features have a significant percentage of null values, such as counts of large firms with over 500 employees which are only present in a few LSOAs. Since the null values represented counts of zero in other areas, nulls were replaced with zeroes in the dataset.

3. Exploratory Data Analysis

The final cleaned dataset has 4765 LSOAs. Now we will look at the descriptive statistics, with visualizations, and make some observations. We will also check feature distributions and correlations.

3.1 Descriptive statistics

The classes in the dataset are highly imbalanced. Of the 4765 LSOAs, only 93 (2%) are located inside the congestion zone.

3.2 Correlations with congestion charge indicator

We expected to see the congestion charge zone indicator have a stronger correlation with features that are more prevalent inside the zone, for example, higher office rents, and weaker correlation with features that are less prevalent, like warehouse counts.

In the results summary table below we can observe the changes that occurred over time within the congestion charge zone, compared to other areas of London. The 'percent change' column shows the features that were positively and negatively correlated with location in the zone.

In summary:

The residential population grew very slightly

The number of jobs grew

Correlations with CC indicator	2001	2011	Pct Chg
Population	-0.046	.0017	0.039
Jobs	0.22	0.25	0.14
Car travel time		0.024	0.024
PT travel time	-0.053	-0.092	-0.045
Office rent	0.27	0.38	0.23
Retail rent	0.2	0.23	0.071
Warehouse rent	0.051	0.091	0.12
Large firms	0.46	0.42	-0.38
Medium firms	0.45	0.47	0.25

Small firms	0.38	0.38	0.16
Micro firms	0.37	0.37	-0.0012
Nonmicro firms	0.64	0.58	0.1
INDUSTRIES			
AFM	0.37	0.38	-0.13
Business support	0.35	0.37	0.072
ComTelRD	0.35	0.36	0.05
Creative	0.4	0.38	0.04
Culture	0.44	0.43	0.15
Development	0.37	0.38	0.12
Education Health Social	0.43	0.2	0.11
Management	0.24	26	0.12
Public utilties	0.28	0.2	0.02
Retail	0.39	0.36	0.06
Transportation	0.27	0.23	-0.055
Wholesale	0.34	0.32	0.11

We also expect to see correlation between features that may overlap or be co-located in the same LSOAs, for example, higher office and retail rents, or higher counts of management consulting and business support firms

We expect to see strong correlation between the same features in different years, for example, job counts for 2001 and 2011, versus percent change

Some features are similar and have potential to be strongly correlated, so the aim is to select those with the most normal distribution and lowest amount of missing data. For example, the counts of firms by size (small, medium, large) versus aggregated together (nonmicro)

Now we will look at the descriptive statistics, with visualizations, and make some observations.

3.3 Summary of Observations

The final cleaned dataset has 4765 LSOAs. Now we will look at the descriptive statistics, with visualizations, and make some observations.

4. Modeling

4.1 Techniques and assumptions

The modeling techniques selected for this project were binary classification models. The project tested several classifiers in an attempt to identify the best one, including logistic regression, k-nearest neighbors, decision trees, support vector machines, and a neural network.

This project compared several binary classifiers and a neural network. GridSearch was used for feature engineering. Models were evaluated using mean squared error, accuracy and precision scores. Model parameters were tuned to improve both.

Imbalanced target class

The highly imbalanced target class calls for classification techniques used for similar problems like fraud or spam detection. In these cases where the target class represents only 2% or less of cases, the paucity of positive cases makes it difficult for the model to learn their characteristics. Therefore techniques that give the model more exposure to the positive target cases are required, such as:

- Balancing class weights. The model parameter class_weight is set to “balanced,” meaning the classes will be weighted in the loss function inversely proportional to their frequency in the dataset. The estimator works to minimize the error on the more heavily weighted positive target class.
- Oversampling from the target class. The target feature is sampled to create a “balanced” dataset with 50% positive and negative cases, prior to running the estimator. In this method the negative cases are undersampled.

Sparse matrix

This dataset has a large percentage of empty cells, it is what is called a “sparse matrix”. In most cases, the null values reflect where there are no firms of a certain class, for example wholesale, or fewer than 10. Filling in null values with imputed values would have reduced the veracity of the data, as null values represented. Therefore they were filled with zeros. Sometimes features with a high percentage of null values are dropped from the dataset, however in this case they were considered important because it is not a very large dataset and the low counts often represent important minority classes, such as counts of large employers.

Validation

The training model will be evaluated using K-Fold cross-validation. To ensure that each class is proportionally represented, we will use Stratified K-Fold cross-validation .

Evaluation Metrics

The confusion matrix was used to compare true to predicted outcomes for the target feature. Since the priority is identifying Positive cases, models will be scored using the True Positive rate, or Recall. Since scoring for Recall will tend to result in more false positives, we will also use the False Negative rate as an indicator of model quality.

The AUC (Area Under the Curve) score is also used as a metric of overall model quality. The AUC measures the area under the ROC (Receiver Operating Characteristics) curve of probabilities between classes. It measures the model’s ability to distinguish between classes, where AUC=1 represents 100% chance the model can accurately predict positive or negative class and AUC=.5 represents a 50% chance of accurate distinction between classes.

The best model will have the highest Recall rate, lowest False Negative rate, and highest AUC score.

Assumptions

An improvement in classifier predictive power over time, as evidenced by higher metrics, means that the congestion charge zone has become more differentiated from other LSOAs. Therefore the same models were run on the three datasets representing the same areas at different points in time and change over time. Results were compared between datasets.

4.2 Feature Engineering

All the predictor features are numeric and had skewed (non-Gaussian) distributions and were therefore normalized (re-scaled) using StandardScalar.

4.3 Feature Selection

As a result, the final dataset used for modeling had 45 features and no missing values.

The most statistically significant features were selecting using the Select Percentile strategy, 10th percentile

4.4 Model test design

Training and validation sets were created for each of the three datasets, with a 30% test ratio.

4.5 Preliminary Models

Two methods of correcting for an imbalanced target class were compared to a baseline Random Forest Classifier model using default settings. Each model was scored using stratified k-fold validation, and GridSearchCV was used for hyperparameter tuning.

Model 1. Random Forest Baseline, no correction for imbalanced classes.

Model 2. Random Forest with Oversampling

Model 3. Random Forest with Class Weights

Results

4.5 Model Improvement

Feature selection and engineering and parameter hypertuning using GridSearch

5. Evaluation

5.1 Final models

Both Models 1 and 3 used the full dataset of 45 features and had very similar results. The alpha value for the Ridge regression in Model 3 was .001, indicating that very little weighting was applied to the model coefficients. It could not improve much, if at all, upon the OLS, or standard linear regression, used in Model 1. Both of these models had very good performance, with a mean squared error (MSE) of .189 and R-squared of .637 for the test set. The low MSE value of is indicates that the model's errors were close to zero. The R-squared indicates the model can explain about 64% of the variability in price, with the remainder due to factors not included in the model.

Model 2 used a selection technique to evaluate each of the 45 features and select only the most statistically significant ones for inclusion in a linear regression model. A threshold of the tenth percentile of significance was used, resulting in the selection of 7 features. The evaluation metrics for this model were not as good as the others. However, the R-squared value of .531 on the test set revealed that these 7 features explained more than half of the sales price.

5.2 Results for Research

Highly explanatory features contributing to increased price were pickup and 8 cylinder engine, and those contributing to a lower price were front wheel drive and vehicle age. The features four wheel drive, condition and sedan were also selected, and all had a weak negative impact on price in this model. Since we expect four wheel drive and better condition to have a positive impact on price, that indicates there is some issue with the data, most likely because both of these variables had a significant proportion of nulls.

All the coefficient values in Models 1 and 3 were of the expected sign. These models also identified pickups as strongly influencing higher prices and vehicle age strongly influencing lower prices. Other factors strongly associated with higher prices were 12 cylinder engines, luxury manufacturers, and Ram, GMC, Jeep and Toyota makes. In general prices increased with engine size. A weaker but positive association on price was due to four wheel drive, convertible type, better condition, and Ford, Honda, Chevrolet makes.

Factors strongly associated with lower prices were vehicle age, odometer mileage, hatchback type and Mitsubishi, Saturn, Mercury, Fiat makes. Age had a stronger influence

on lower prices than mileage, indicating that customers value a newer car over higher mileage. Factors weakly associated with lower prices were forward sedan or wagon types, 4 or 6 cylinder engines, forward drive, and Chysler, Nissan, Hyundai, and Kia makes.

Summary of Model Evaluation Metrics

	Mean Squared Error	R-squared
Baseline	Train MSE: 0.519 Test MSE: 0.524	
Model 1	Train MSE: 0.185 Test MSE: 0.189	Train R-Squared: 0.643 Test R-Squared: 0.637
Model 2	Train MSE: 0.241 Test MSE: 0.245	Train R-Squared: 0.534 Test R-Squared: 0.531
Model 3	Train MSE: 0.185 Test MSE: 0.189	Train R-Squared: 0.643 Test R-Squared: 0.637

Summary of Coefficients

	Models 1 & 3	Model 2
++	12 cylinders Luxury Pickup Ram, GMC, Jeep, Toyota	Pickup 8 cylinders
+	10 cylinders Convertible 4WD Ford, Honda, Chevrolet Condition	
-	Wagon, Sedan Chysler, Nissan, Hyundai, Kia 4 or 6 cylinders FWD	4WD Condition Sedan
--	Odometer Hatchback Mitsubishi, Saturn, Mercury, Fiat Age	FWD Age

4. Evaluation

The descriptive statistics and linear regression model estimated from the dataset have resulted in useful insights.

4.1 Assessment of Data Mining Results

The data mining goals are to a) identify the most valuable attributes of used cars, and b) predict the sales price of a vehicle. The first goal has been achieved, as shown in the section above. The second goal has also been achieved, as Model 1 could be used to predict a sales price with 64% accuracy.

The success criteria of a pricing model which achieves a minimum of 80% accuracy was not achieved. One limitation was the dataset itself. Although it seemed robust with nearly 500,000 listings, after data cleaning we were left with only 150,000. Many of those listings were missing important nature such as vehicle condition. The models developed here could be improved with a larger and more complete dataset.

4.2 Review of Business Questions

This analysis have achieved the primary client objective of identifying the primary drivers of used car value. Based on the attributes identified, the client should be able to quickly assess a new vehicle when it arrives and calculate an attractive sales price using the pricing model.

5. Deployment

The final model and all documentation may be found at this link.