Using Urban Data Science to Predict Property Development Planning Successes: Recommendations to Lumiere Property

Inessa Tregubova
Chen Xu
Mohd Sarim
Reka Vonnak

Submitted in partial fulfilment of the requirements of
Urban Analytics Group Project, University of Glasgow

2020

Table of Contents

1.Introduction *3*

2. Literature Review *3*

3. Data *4*

3.1 Data Collection **4**

3.2 Feature Engineering **6**

3.2.1 Application features 6

3.2.2 Proposed and land registry areas 6

3.2.3. Density and height 6

3.2.4 IMD Data 8

4. Exploratory Data Analysis *8*

4.1 Bromley **8**

4.2 Comparison to Croydon **11**

5. Methods *15*

5.1 Datasets **15**

5.2 Logistic regression **16**

5.3 Support Vector Machines (SVM) **17**

5.4 Xgboost **17**

6. Results *19*

6.1 Results for Bromley **20**

6.1.1. Logistic Regression 20

6.1. 2 SVM 21

6.1.3 Xgboost/ Decision tree 22

6.2 Results for Croydon and Bromley **24**

6.2.1 Logistic Regression 24

6.2.2 SVM 24

6.2.3 Xgboost 25

7.Discussion *26*

7.1 Conclusion **26**

7.2 Recommendations **26**

7.3 Limitations **27**

7.4 Future work **27**

8. References *28*

Using Urban Data Science to Predict Property Development Planning Successes: Recommendations to Lumiere Property

1.Introduction

Property development is subject to various expectations and regulations, both with regards to customers and planners. Property developers operate in this intertwined market, where even the smallest detail is decisive. Investing into a project requires both significant financial and human costs, and only successful applications yield profit for the company. This report is the outcome of a three-month long collaborative project with Lumiere Property to explore the factors that make planning applications successful in a given area, aiming to help them plan more successful applications.

Lumiere Property is a property development company specialising in residential properties. Their approach to selecting suitable development sites include an automated algorithm that encodes the official planning guidelines. However, even with different algorithms they face the challenge that the list of rules provided is not exhaustive. In many cases it is not known on what grounds planning officers decide on a proposed development, making it financially risky for smaller property developers to submit an application. As guidelines can vary between different councils it is important to understand what makes planning applications successful in a particular locality. Bromley, a Greater London borough was chosen for this analysis, as overall planning success rates in the borough fluctuate between 50 and 55%. The aim of the collaboration was to use urban data science to build a model that predicts successful applications in Bromley, making it a collaboration of academic knowledge and professional expertise. Moreover, this new approach of estimating the probability of successful planning applications can contribute towards a more efficient urban renewal approach. This report sums up the work that went into this project and provides a recommendation based on the results.

The rest of this work is organised as follows. Section 2 examines some relevant literature. In section 3 we detail the available data and the additional data collection methods. Section 4 compares the two boroughs, while sections 5 and 6 discuss methodology and results. Section 7 offers some conclusive results and recommendations to the company.

2. Literature Review

Individual property renewal is often seen as a ‘free market renewal approach’ (Porat and Shach-Pinsly, 2019). The primary reason of renewal is improving the existing properties, mostly by

demolishing old ones and building a new one on the existing area (Carmon, 2001 ; Kleinhans, 2004). It presents opportunities for unlocking potential values as well as improving the community living there. However, both time and costs associated with acquiring planning permission and undertaking specialised assessments – such as evaluating the environmental impact – are major issues in planning applications (Brown-Luthango, Makanga and Smit, 2012).

There are two related perspectives when it comes to predicting the success of planning applications; the physical construction and demographic characteristics. Primarily, planning applications need to follow Planning Practice Guidance to make sure the buildings are safe and promote local sustainable growth. Also, it needs to consider the effect of local neighbours and local characteristics. (Chan and Liu, 2018) show that neighbourhood building density and neighbourhood building height are significantly associated with occupant health. The reasonable visibility of roads is also affected by the surrounding buildings. Additionally, the influence of socio-economic characteristics on housing demand is also considered, as the planning applications need to meet the needs of local residents. Residential construction depends on the population's age structure, household, and the household's human capital (Geyer, 2017; Eichholtz and Lindenthal, 2014; Lindh and Malmberg, 2006). In some cases, the education level, health conditions, and income levels are positive correlated with the housing demands of residents (Eichholtz and Lindenthal, 2014). In this report, housing demand will be reflected by the number of bedrooms.

Based on the literature, this study takes building density, building heights, and the Index of Multiple Deprivations (IMD) as additional features. These will be detailed subsequently.

3. Data

3.1 Data Collection

The Bromley planning applications dataset provided by Lumiere property has only 113 unique observations. This could greatly limit the performance of machine learning algorithms used later in the report. Upon further discussions, it was suggested that planning application data from the London borough of Croydon could help increase the data points and make the model robust. Located just next to Bromley, Croydon was selected due to its similarities with Bromley. A list of Planning Applications for Croydon was shared with us by Lumiere Property. The process to obtain the data was also explained. The steps are summarised below.

Firstly, the land registry polygon shapefile was loaded into ArcMap. Next, a shapefile with all the required attributes was created. The attributes contained the maximum height, parking, number of units of each type (1,2,3, or 4+ bedrooms or studio), whether the property was on a corner, the status of the application and reasons for rejection, if any. Then, the application number was used to get the application details from the Croydon Planning Portal. The proposed site plans for the building were then geo-referenced to the land-registry polygons, as shown in Figure 1.

Figure 1

Once the site plan was placed on the polygon, the editor tool was then used to mark the built area polygon. The Create Features option in the Editor toolbar was used to add new features for the Croydon applications, and all the attribute values were entered based on the data from the applications portal.

Figure 2

The edits were then saved and the Stop Editing option was used to save the values. A total of 101 polygons were created like this. Whenever there were multiple buildings in the same proposed area, separate polygons were created for each building and added to the attribute table. Thus, a total of 94 planning applications were coded for Croydon.

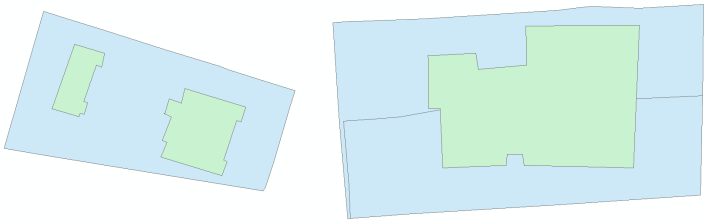
3.2 Feature Engineering

3.2.1 Application features

Parking space is one of the common reasons for rejection. We created “parking per unit” feature to show how many parking spaces are available for each unit. Since there are four features related to number of bedrooms, we sum up them and create two new variables: “Two more bedrooms”, “Three more bedrooms”. We also calculated average volume of each unit within one proposed building, which is multiplied height by area and divided by the number of units.

3.2.2 Proposed and land registry areas

On a closer look we found that size is the main rejection reason in Bromley. To examine this, we calculated the area of proposed applications and spatially joined it to the land registry area. We downloaded land registry map from the UK Government Website. In some cases, there are multiple proposed buildings on one land registry polygon. These were summed up together. In the same way, when one proposed building extends over multiple land areas, they are also summed up together (see Figure 3 and 4). Nine planning applications in Bromley and one planning application in Croydon could not be matched to the relevant land area in the official registry. These were removed from the analysis.



Figures 3 and 4

3.2.3. Density and height

Building footprints data (2017) from Digimap was used to calculate the height of neighbours and the density of buildings in Bromley and Croydon. The planning application shapefiles were

spatially joined with the data zones shapefile to get the data zone for each planning application. The geographic unit for these is the Lower Layer Super Output Area (LSOA).

As discussed in the literature review, size and local effects are important factors in the decision-making process. Therefore, the height of the proposed building was compared to the height of its neighbours. To define 'neighbours', a 40 meters buffer was created around the proposed building (Figure 5). For each proposed building, we calculated the maximum height and medium height of its neighbours, considering properties with an area at least 80% of the smallest area of proposed buildings.

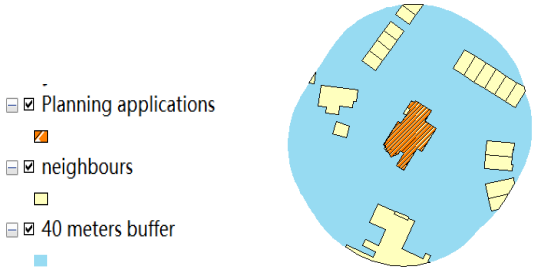


Figure 5

Population density and building density was also computed. This is an important measure for new developments as it can signal the area's need for housing. The population density data is from the 2011 Census. This information was registered based on the LSOA area as well. Two measures of built density were defined. First, building density as the ratio of proposed building area to the whole land parcel. Then, the building density of the entire neighbourhood was calculated by summing up the total built-up area of the LSOA and then dividing this number by the total land area (Figure 6).

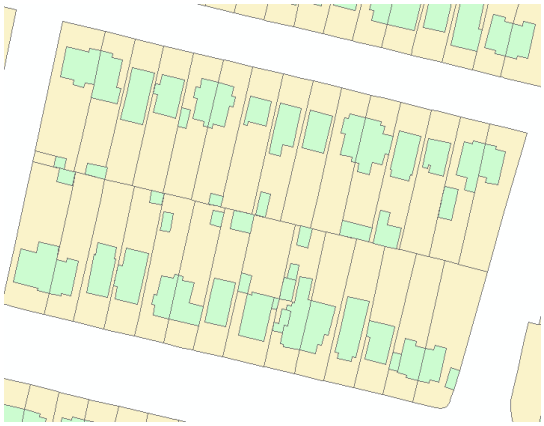


Figure 6

3.2.4 IMD Data

The Index of Multiple Deprivation (IMD) data was also included in the analysis. It allows us to understand the socio-economic characteristics of the area under study by providing scores on the relative levels of deprivation in each area. There are fifteen sub-domains of deprivation listed in the table below.

Table 1 - IMD data

Index of Multiple Deprivation (IMD) Score	Income Score	Employment Score	Education, Skills and Training Score
Health Deprivation and Disability Score	Crime Score	Barriers to Housing and Services Score	Living Environment Score
Income Deprivation Affecting Children Index (IDACI) Score	Income Deprivation Affecting Older People (IDAOPI) Score	Children and Young People Sub-domain Score	Adult Skills Sub-domain Score
Geographical Barriers Sub-domain Score	Wider Barriers Sub-domain Score	Indoors Sub-domain Score	Indoors Sub-domain Score

4. Exploratory Data Analysis

4.1 Bromley

Bromley – a borough in South-East London – is the focus of this study. Here we provide some basic information about the area. It is made up of various smaller towns and has a population of around 300,000 people (ONS). It is seen as one the most affluent areas in Greater London with a higher percentage of owner occupiers than other areas in Greater London. According to the 2011 Census, Bromley's population is socio-economically homogeneous with a high percentage of UK-born residents.

Below are a few graphs which explore the Bromley dataset. First is the visual illustration of the outcome of the planning applications (Figure 7). 47% of applications were approved and 53% rejected. The two most common rejection reasons are size and neighbourhood effects.

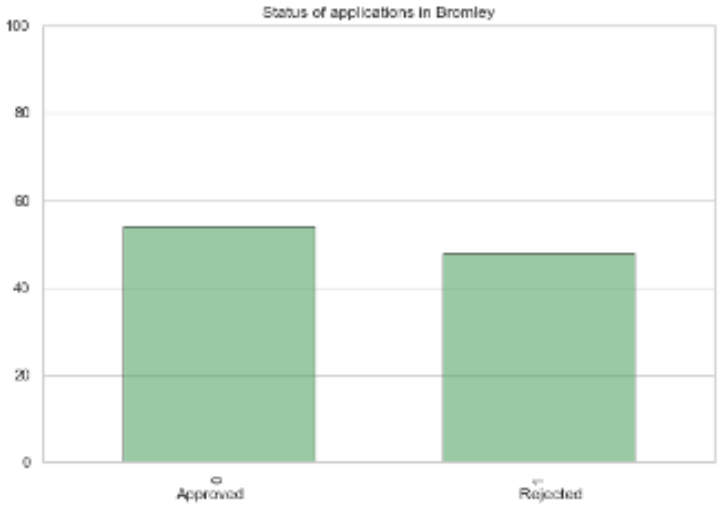


Figure 7

Next is height and built density. On the left the side of Figure 8 is the maximum height of applications in Bromley with blue and the maximum height of the neighbouring properties in green. The applications in the dataset tend to be smaller than neighbouring houses. On the right we can see the ratio of applications' heights to that of their neighbours', which is a more informative measure than simple height. We can see that in a few cases the applications are higher than their neighbours (within the 40-meter buffer).

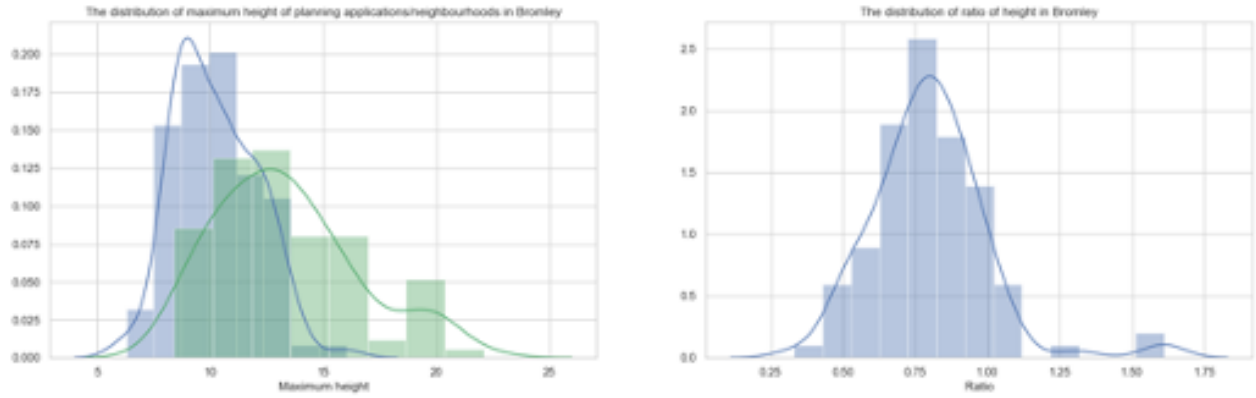


Figure 8

However, to see whether taller buildings were more likely to be approved we can look at Figure 9. The x-axis shows the buildings height in meters, while the y-axis is the ratio of application height and the median height of neighbours. Approved applications are shown in yellow and rejected ones in blue. In quite a few cases tall buildings were approved, an interesting finding given the council's insistence on neighbourhood similarity. This will be explored in more detail later.

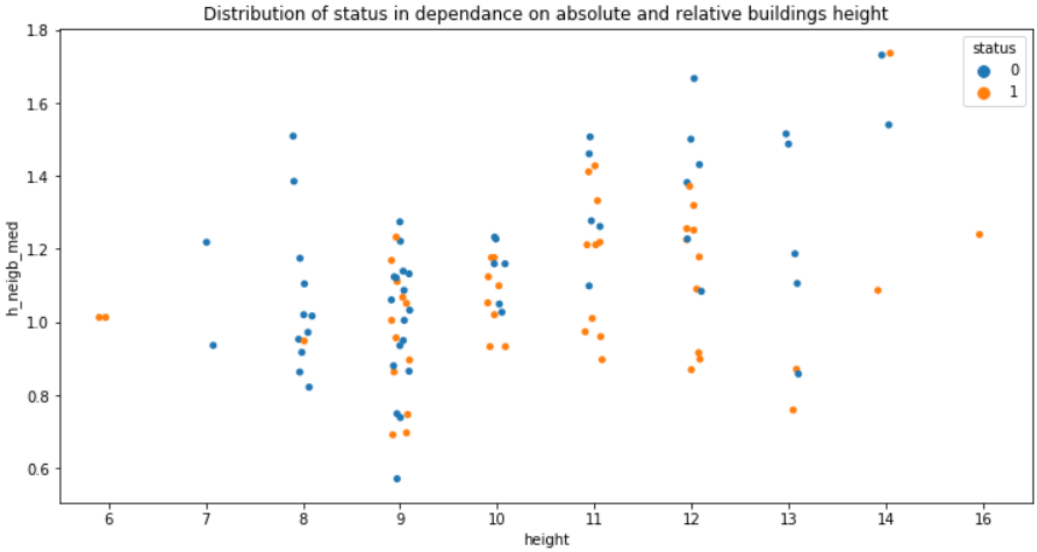


Figure 9

Density was explored in two different ways: built and population density. Figure 10 shows the number of residents per hectare and the built area ratio. Since Bromley is a family-friendly area, built density is relatively small (5-20%), leaving a lot of green space empty.

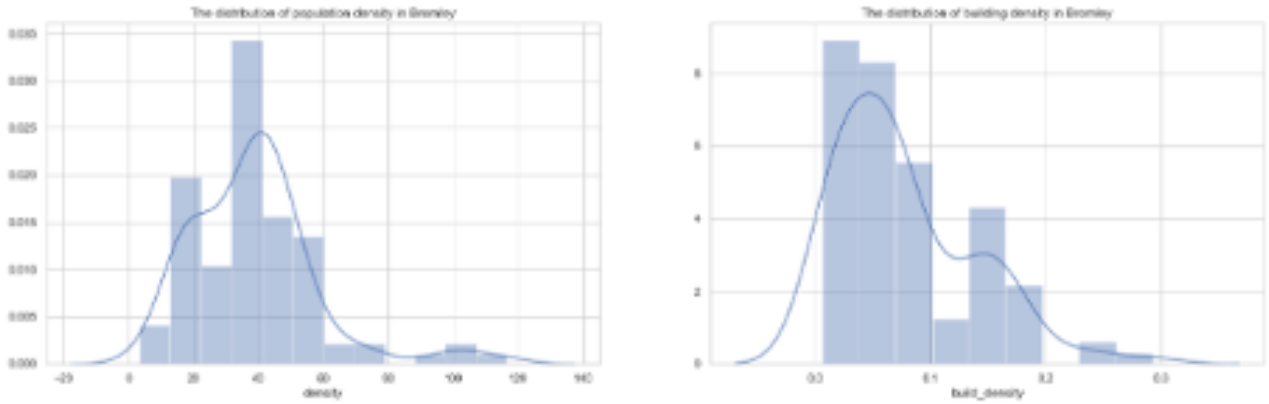


Figure 10

The next important thing to look is the availability of parking spaces for each application. As it will be detailed, parking space was the fourth most commonly cited reason for rejection in the Bromley dataset. The left side of Figure 11 shows the cumulative number of parking spaces per application. Since this can vary based on the number of units in the application, the right plot shows the number of available parking spaces per unit. Most applications provide at least one, if not two spaces. This can ensure that roads will not be overcrowded if a new housing unit is being built, which is a priority for planning officials.

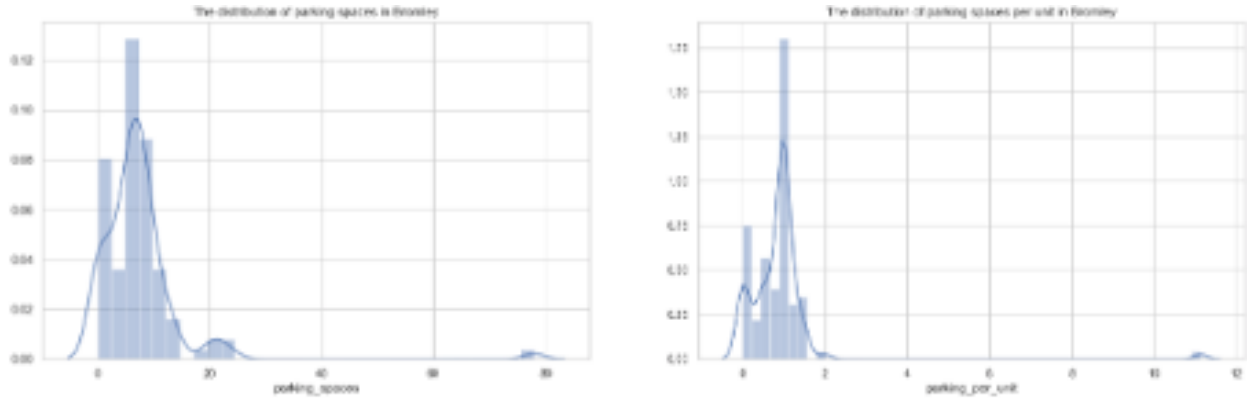


Figure 11

4.2 Comparison to Croydon

Croydon is the borough that was chosen for complementary analysis. It is part of Greater London and has the highest population of all boroughs (around 360,000 residents as of 2011) (ONS). While it has similar demographic, economic, and social characteristics to other neighbouring areas, there is a larger share of children and young adults.

The Croydon dataset contains 94 observations, each of them coded by the authors. It has a lower rejection rate of planning applications. As seen in Figure 12, 78 planning applications were approved and only 16 rejected.

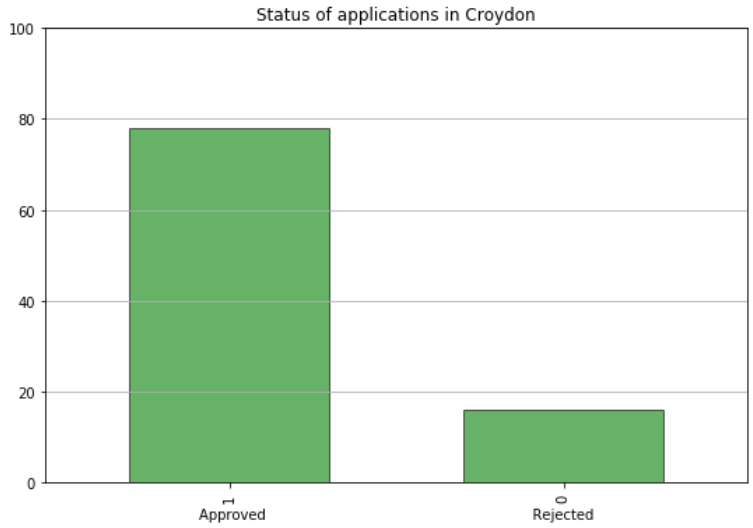


Figure 12

Compared to Bromley, Croydon has multiple reasons for rejection. Local characteristics, parking spaces, and neighbourhood effects are commonly cited reasons for rejection (Figure 13). Neighbourhood effect seem to be a problem only in a third of rejected applications in Bromley, but in nearly half in Croydon. In Bromley, local character is the least important criteria, whereas in

Croydon it is a more common reason. These differences suggest that the outcome of an application is explained by different factors in the two areas.

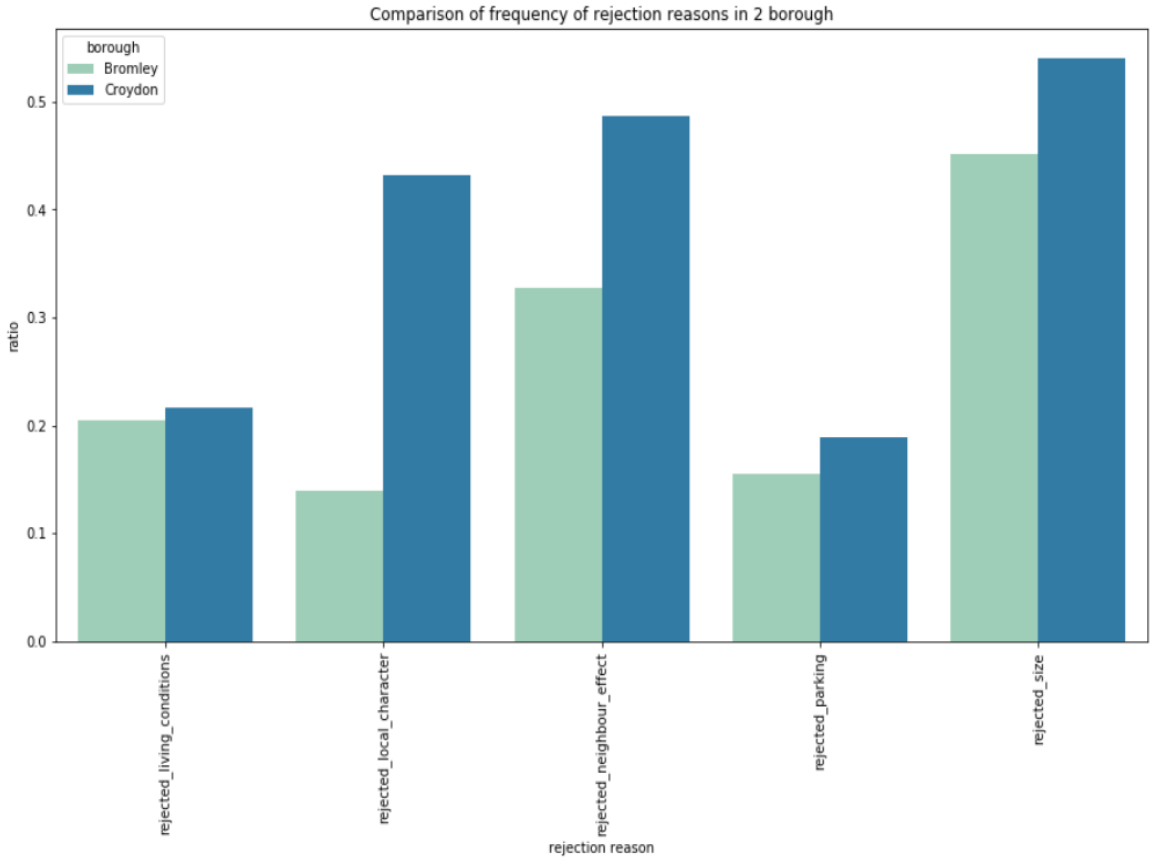


Figure 13

To follow up 'size' as the rejection reason we next look at height. In Croydon there is a bigger variation of height ratio (between 0.5 and 1.5 of neighbours), suggesting that significant part of applications exceed neighbour's height. In contrast, 90% of observations are shorter than their neighbours in Bromley.

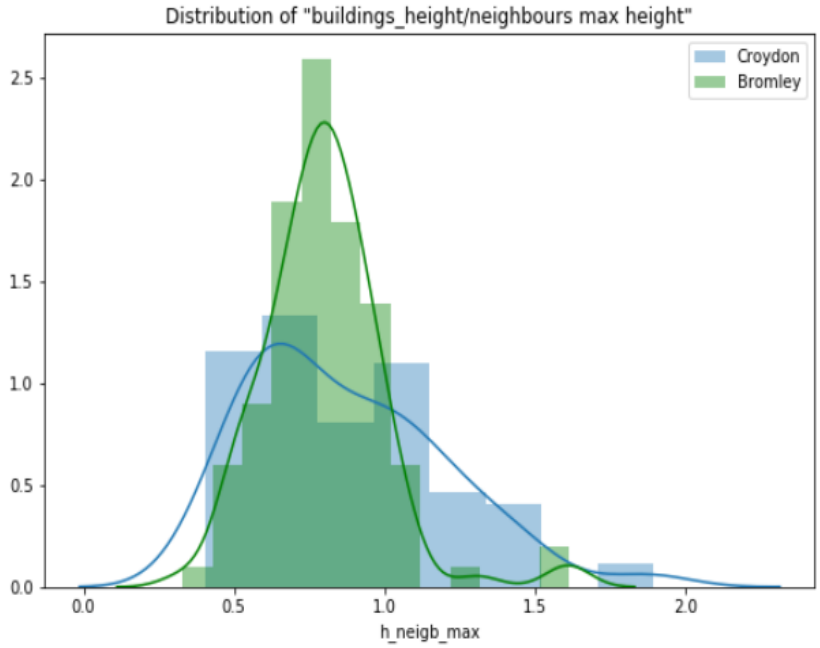


Figure 14

Moreover, proposals in Croydon have more units per land area than Bromley. The average number of units is 10 in Croydon, while most values fall between 6 and 8 in Bromley.

Next is the type of units. Two or more bedrooms are more family-friendly than either one-bedrooms or larger, four+ bedrooms. This difference between the two boroughs reflect housing needs. As we could see from the description of Bromley, it is more of a traditional suburban area with slightly older residents and families. Croydon on the other hand has a higher share of young adults and professionals, who could prefer to or could only afford to live in one-bed units.

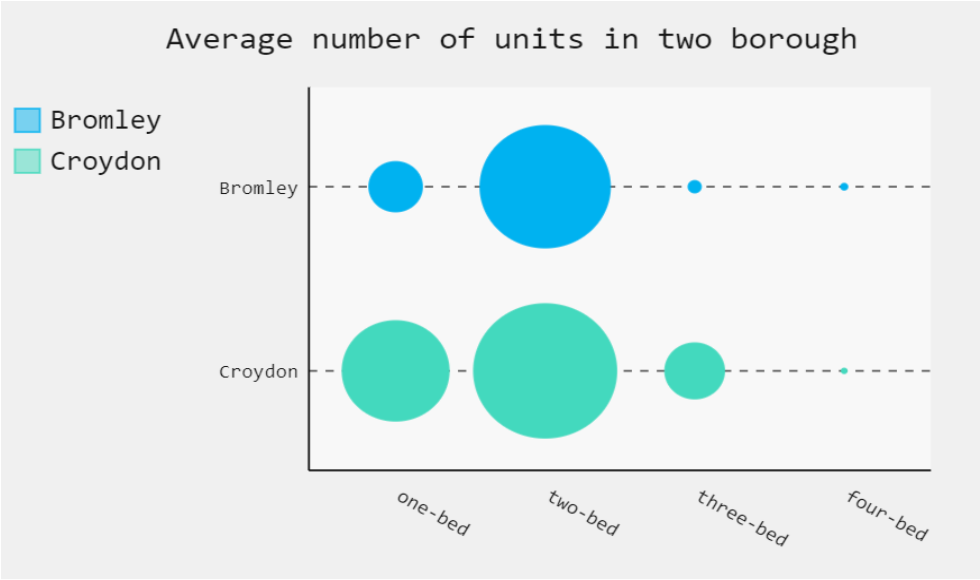


Figure 15

In terms of built area, our expectation is that if a proposed building covers significant part of the land it is more likely to get rejected. The boxplot on Figure 16 seems to confirm this hypothesis,

as rejected plans have a higher built area-land ratio. Building density is higher in Croydon (ranging between 0 and 36%) compared to Bromley (up to 20%).

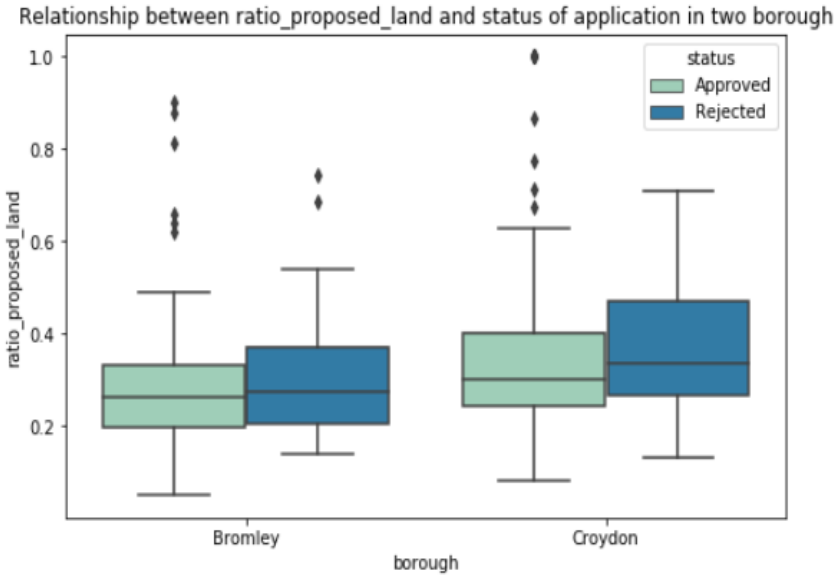


Figure 16

Parking is also an important factor to consider, as it is a common reason for rejection. In Bromley they proposed at least one parking space per unit, whereas in Croydon the average is one for two units (Figure 17). This suggests that some residents need to rely on public parking spaces, which can affect traffic and pedestrian access to pavements.

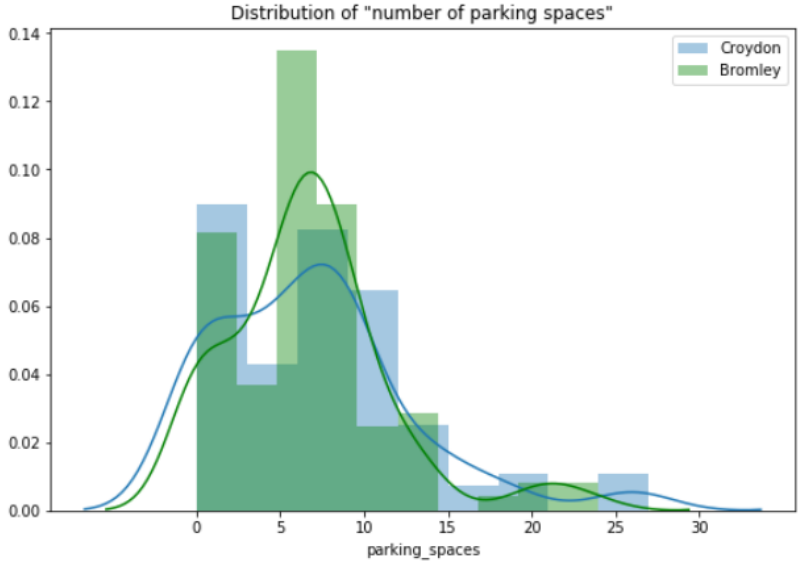


Figure 17

In order to sum up the differences between the two boroughs we calculated the correlation of features to the target value – the status of the application – for each dataset (Figure 18).

The diagram below indicates that in the most cases the correlation has the opposite directions in the two datasets. However, more data is needed to conclude any message from this plot, as the correlation coefficients are relatively low.

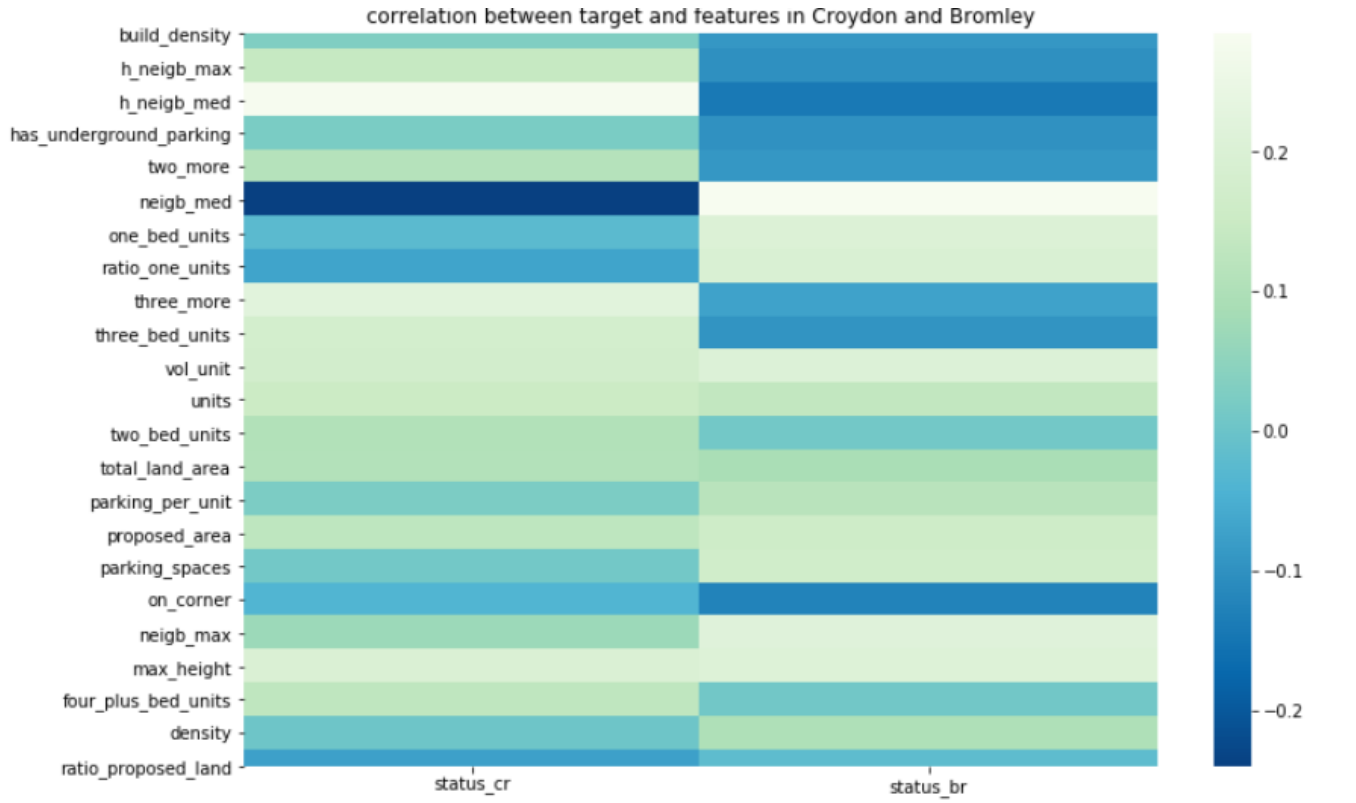


Figure 18

5. Methods

5.1 Datasets

Table 2 shows how the dataset was split into training and test samples for modelling. Table 2 contains values for the combined dataset of Bromley and Croydon.

Table 2 Bromley

	Train sample	**Test sample**
Number of observations	81	21
Number of approved applications	38	10

Target variable	Planning application was approved either directly or after appealing	
Approved applications	46.9%	47.6%

Table 3 Bromley and Croydon

	Train sample	Test sample
Number of observations	156	39
Number of approved applications	100	25
Target variable	**Planning application was approved either directly or after appealing**	
Approved applications	64.1%	64.1%

The target variable in this project is the outcome of the planning application, denoted by the variable 'status'. Approval is coded as '1' and Rejection is '0'.

5.2 Logistic regression

Logistic Regression is a classification model that belongs to the family of Generalized Linear Models. The main advantage of Logistic Regression over Linear Regression is that it is robust to outliers due to the use of the sigmoid function. It predicts the probability of success and then uses the sigmoid function to classify the probabilities into categories. The coefficients for the training data are estimated using Maximum Likelihood Estimation (MLE).

There are many advantages of Logistic Regression. First, it is easy to implement and train, and the coefficients can also be interpreted. Second, the features do not need to be scaled. However, there are also some disadvantages, namely the assumption of linearity and need for selection of relevant features for good performance.

Grid search was used to find the optimal hyperparameters for the model. It requires an estimator and a grid of parameter values from which the best performing parameters are selected. In this case, the estimator of GridSearchCV was Logistic Regression. The range for the parameter C was

specified between 0.001 and 1000, while the penalty parameter was chosen from L1 (Lasso) and L2 (Ridge). The GridSearchCV library from sklearn.model_selection was used in this case. This technique combines GridSearch with K-Fold Cross-Validation to find the best parameters for the training set. Here, the dataset is divided into K parts and the model is trained K times with one of the parts as the test set and the remaining as training sets. The final result is the average of all the cross-validation results.

5.3 Support Vector Machines (SVM)

SVM is a popular Machine Learning algorithm for both regression and classification. It was chosen as a method to model planning applications, as it works relatively well on smaller datasets. One reason behind SVM's popularity is the so-called 'Kernel-trick', which makes it possible to gain the same results as if computing in multiple dimensions, without actually doing so. However, since it calculates the dimensions mathematically the computation time tends to be more compared to other models.

The important parameters that need to be hypertuned are 'C' (which is responsible for regularisation), the gamma (which controls overfitting), and the type of kernel. The most often used kernels are linear, polynomial, and RBF (Géron, 2019).

Just as in the case of Logistic Regression, we used GridSearchCV to hypertune the parameters and validate using K-fold Cross-Validation. However, there are some disadvantages of SVM, like the fact that they are sensitive to scales (so a scaler needs to be used beforehand) and that they do not output probabilities for each class.

5.4 Xgboost

XGBoost is an optimised distributed gradient boosting library designed to be highly efficient and flexible (Xgboost Developers, 2020). It implements Gradient Boosting machine learning algorithms. XGBoost provides a parallel tree boosting. The advantage of this algorithm over simple gradient boosting is the inclusion of a regularisation parameter. The regularisation term controls the complexity of the model, which helps to avoid overfitting and better generalise results. Therefore, it can be used on small datasets.

To address outliers, we used *GridSearchCV* which performs cross-validated search over a parameter grid and finds the best combination of parameters. The final score for the best combinations of parameters is 0.93 on train sample and 0.76 on test sample. 27 features were marked as important that is a big number for such a small dataset. These are signs of overfitting.

Figure 19 below shows how this effect takes place as the test curve is going down once we pass six estimators.

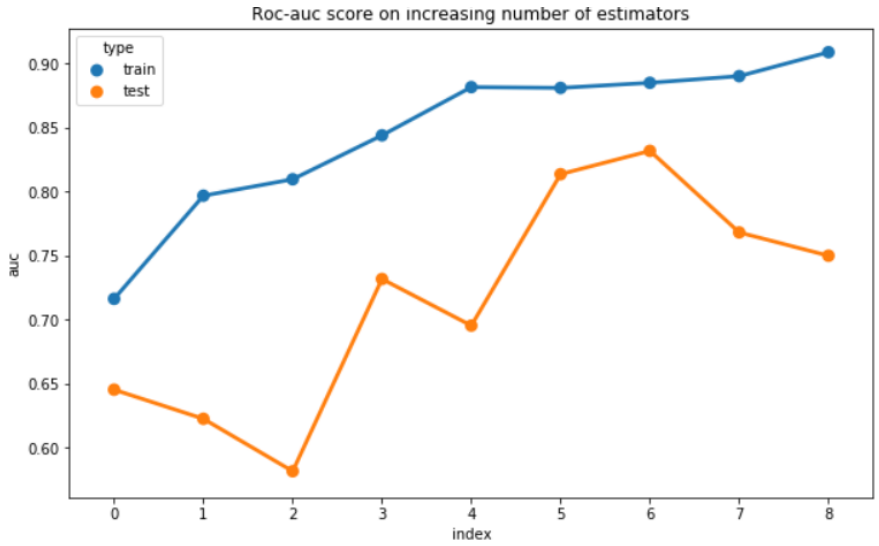


Figure 19

To address this issue, we first removed irrelevant IMD features and then run bootstrap with decision tree. Bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. Unlike cross-validation it estimates the machine learning model skill with confidence intervals.

We set number of iterations equal to 1000 and the percent of random sample to 90%. For the Decision Tree parameter, we used specified max depth as 7 and minimum samples leaf as 2 as we do not want to have a long chain of rules to predict status.

Finally, the distribution of test scores presented on the diagrams below was achieved.

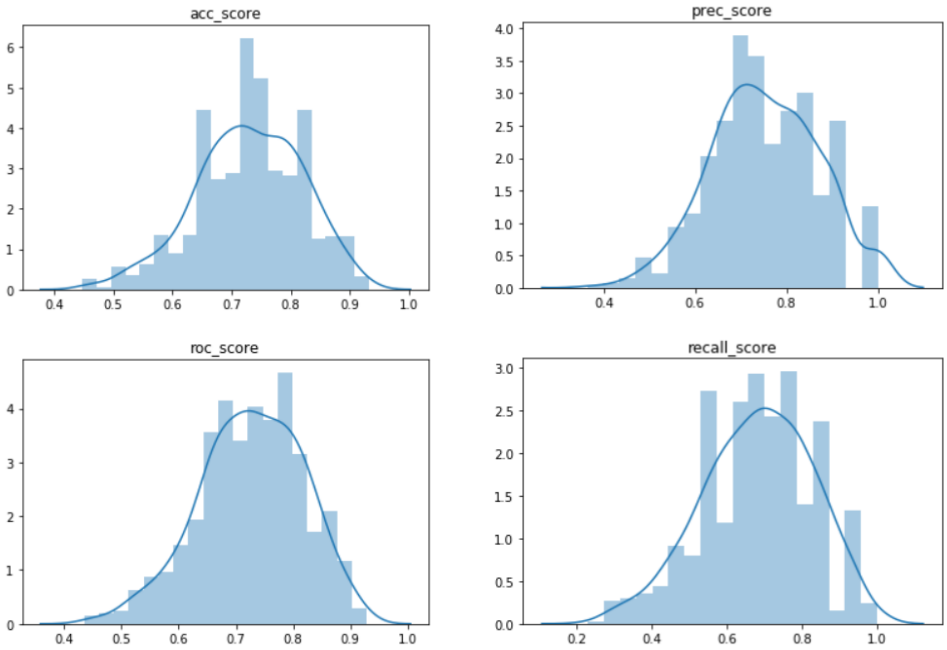


Figure 20

With 95% confidence intervals the ranges of values for test scores are summarised in Table 4. Once we got the range of values for 'real' scores, we built a final decision tree, whose results lie in the defined intervals.

Table 4 – Confidence Interval for bootstrapping

Roc_auc_score	71.7%-73.5%
Accuracy_score	72.1% - 73.8%
Precision_score	74.2% - 76.4%
Recall_score	66.9% - 69.6%

After adding Croydon data to Bromley, the following pipeline was built to predict the status of applications:

- Scale to unit variance using StandardScaler
- Run selectFromModel with lasso regression (Logistic regression with L1-regularisation) to reduce model complexity
- Run GridSearchCV with Xgboost on the variables to predict the target variable

Since combining the two datasets doubled the size of our rows, we can be more confident with the stability of results. We also used logistic Regression model with L1-regularisation to decrease the model's complexity and leave only the most important features. Finally, we run Xgboost with GridSearchCV repeating the steps we have done for Bromley.

No additional manipulations were done on the two datasets as results are just slightly better than a random guess. We can still assume that some features are spuriously correlated with the target variables and rather indicate the borough than status of decision. Thus, the real score would be even worse. Detailed results are provided in the next section.

6. Results

Table 4 and 5 contain the outcome and the ranking for all three models for Bromley only and for the combined datasets.

Table 4 - Test results for Bromley

Model	**Roc-auc**	**Precision**	**Recall**	**Accuracy**
LR	**0.72**	**0.72**	**0.71**	**0.71**

SVM	**0.60**	**0.47**	**0.3**	**0.65**
XGB	**0.75**	**0.77**	**0.67**	**0.75**

Table 5 - Test results for Bromley and Croydon

Model	**Roc-auc**	**Precision**	**Recall**	**Accuracy**
LR	**0.65**	**0.68**	**0.62**	**0.62**
SVM	**0.64**	**0.28**	**0.45**	**0.73**
XGB	**0.52**	**0.65**	**0.76**	**0.59**

6.1 Results for Bromley

6.1.1. Logistic Regression

Since Logistic Regression is sensitive to correlated features, and the dataset has some attributes which are a combination of others. The first step was to ensure mutual exclusion among these features. Due to this, two new attributes were added to the dataset. The attribute 'upto_two_bed' contains the number of one and two bed units, and similarly 'upto_three_bed' contains the number of one, two, and three bed units. Then, the features were divided into mutually exclusive categories as shown in the table below:

Table 6 - Feature Combinations for Logistic Regression

Type of Attribute	**Feature Combinations**			
Bedrooms	'one_bed_units', 'two_more'	'upto_two_bed' 'three_more'	'upto_three_bed', 'four_plus_bed_units'	'one_bed_units', 'two_bed_units', 'three_bed_units', 'four_plus_bed_units'
Area	'proposed_area', 'total_land_area'	'ratio_proposed_land'		
Height	'max_height', 'neigb_max', 'neigb_med'	'h_neigb_max', 'h_neigb_med'		

Only one set of columns were selected at a time from each category of attributes. The data was divided into train and test sets in the ratio 80:20. The train test was used to train different combinations of the features and the results were recorded.

Firstly, only the Bromley dataset was used to train and test the model.

Table 7 –Logistic Regression Results for Bromley

Metric	Train	Test
Roc_auc_score	0.8	0.72
Accuracy_score	0.8	0.71
Precision_score	0.8	0.72
Recall_score	0.8	0.71

The selected features in this case are `('units', 'one_bed_units', 'two_more', 'proposed_area', 'total_land_area', 'max_height', 'neigb_max', 'neigb_med', 'on_corner', 'density', 'build_density', 'Crime Score')`

6.1. 2 SVM

As mentioned, SVM is highly sensitive to scaling, so the MinMaxScaler() from sklearn library was used before splitting the data. The best results were obtained for SVM using GridSearchCV with CV specified as 3 (just like in the case of Logistic Regression and Xgboost). It is important to note that with a 20%-80% split there was a great variation between test runs, as it matters to a large degree how the data was split by the algorithm.

The best result for the Bromley dataset was obtained with the following parameters:

- Best kernel: ’RBF’
- Best CV parameters {'C': 10, 'gamma': 'scale'}
- Best CV accuracy 0.6794871794871796
- Test accuracy of best grid search hypers: 0.65

The summary of results is the following:

- Roc-auc score: 0.604
- Precision: 0.47
- Recall: 0.3

These results can be attributed to the small sample. In the next section we are going to see whether combining the two datasets will make the results more profound.

6.1.3 Xgboost/ Decision tree

Running decision tree with the best parameters found due to applying bootstrap technique gave the 75% of accuracy of prediction with 77% on precision score.

Table 8 – Xgboost results for Bromley

Metric	Train	Test
roc_auc_score	0.93	0.745
accuracy_score	0.93	0.75
precision_score	0.97	0.77
recall_score	0.815	0.67

Among the most important features are the number of units, absolute and relative maximum height of buildings in neighbourhoods, area of proposed building, ratio of one-bed units in building, ratio of building height to land area. All these features refer to the size of building. Thus, we can explain by fact the size is the most common reason for rejection in Bromley.

Additionally, two IMD scores 'Geographical Barriers Sub-Domain Score', 'Barriers to Housing and Services Score' were marked as important. These indicators measure the physical and financial accessibility of housing and local services. Their impact on the target variable points out the importance of the neighbourhood's local characters in the decision-making process.

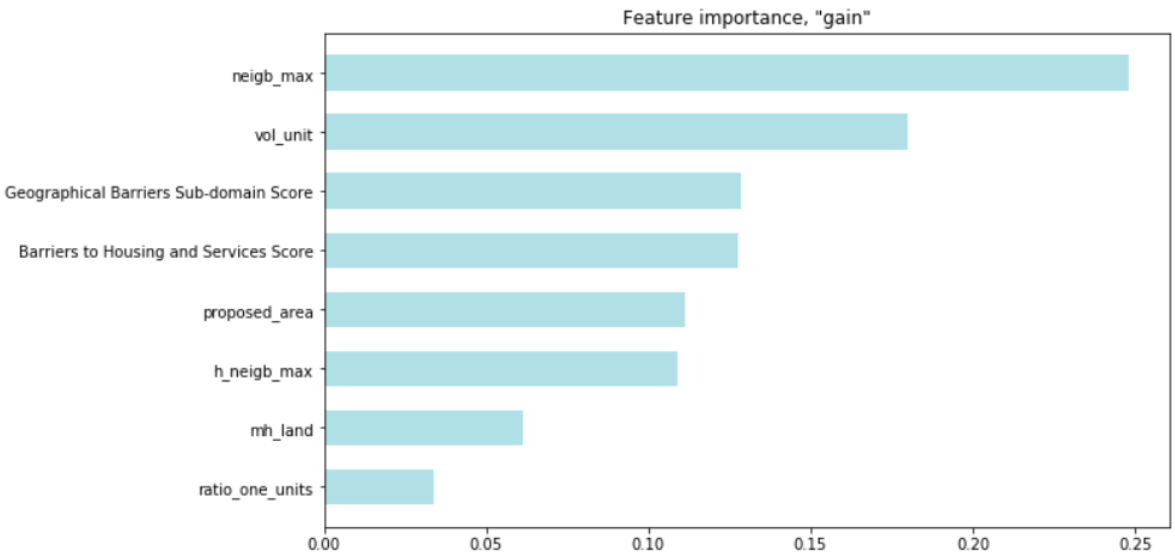


Figure 21

Afterwards we visualised the decision tree to identify the relationships between the most important features and the target variable. According to the chart, decisions on application are more likely to be approved in 'taller' neighbourhoods (with max height bigger than 15 metres while the average height in Bromley is around 8 metres). If the density of units in proposed building is such that every unit has less than 291 of volume (around 50 sq. metres of area) then applications are likely to be rejected. Lastly, in areas with low IMD scores which means better living conditions the level of approval is higher.

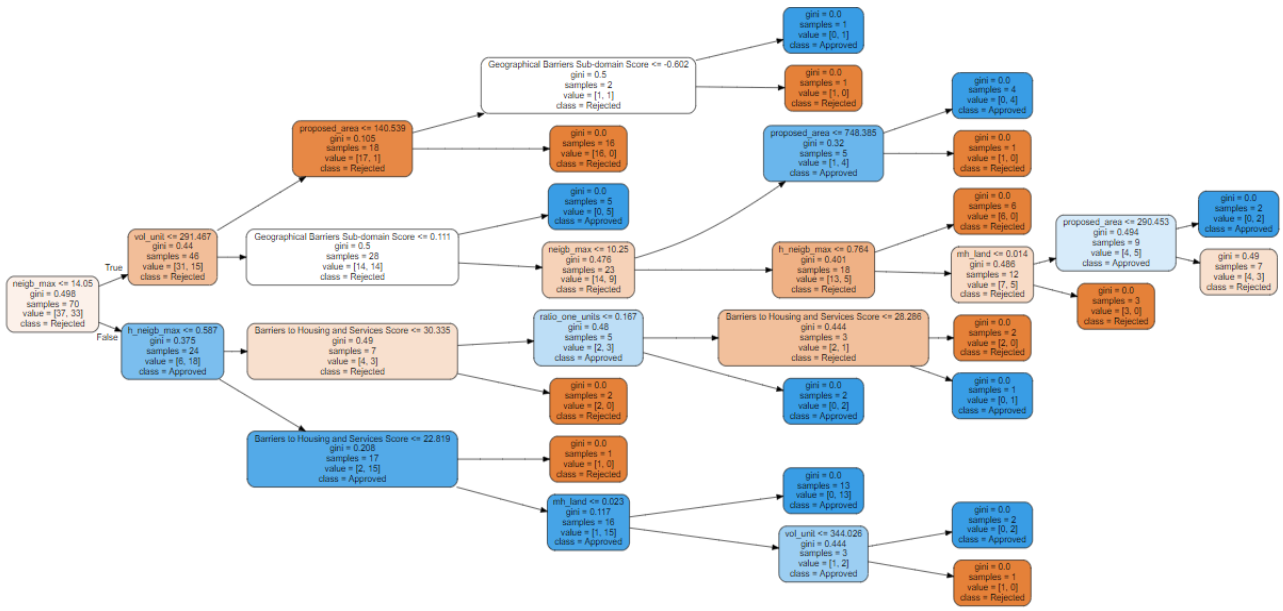


Figure 22

6.2 Results for Croydon and Bromley

6.2.1 Logistic Regression

Next, Croydon dataset was merged with Bromley to determine whether it improves the accuracy of the model. The same procedure was followed as in the case of Bromley (detailed in section 6.1.1).

Table 9 – Logistic Regression Results for Bromley + Croydon

Metric	Train	Test
roc_auc_score	0.58	0.65
accuracy_score	0.56	0.62
precision_score	0.63	0.68
recall_score	0.55	0.62

The selected features in this case are `('units', 'upto_three_bed', 'four_plus_bed_units', 'proposed_area', 'total_land_area', 'max_height', 'neigb_max', 'neigb_med', 'on_corner', 'density', 'build_density', 'Crime Score')`

6.2.2 SVM

The combined dataset combines 196 observations. Similarly in the case of the Bromley dataset alone, the MinMaxScaler() was prior to analysis. To offset the class imbalance the target value was inversed. This means that instead of predicting approved application we predict rejected ones.

The best results with the combined dataset was reached using the following parameters:

- Best kernel: ’RBF’
- Best CV parameters {'C': 1, 'gamma': 1}
- Best CV accuracy 0.6606854838709678
- Test accuracy of best grid search hypers: 0.725

The summary of results is the following:

- Roc-auc score: 0.641
- Precision: 0.275
- Recall: 0.454

While these results might seem like an improvement, it is important to difference between test runs due to the size of the dataset, even if by combining the datasets we essentially doubled the number of rows. It appears that while SVM produces some relatively good results, the computation time and the variance between runs make it a less reliable model for the datasets.

6.2.3 Xgboost

As we expected small number of features in our data can predict decisions for two boroughs together. Due to this the quality of the model is very low on test dataset and roc score is just 2% better than random guess.

Table 10 – Xgboost Results for Bromley + Croydon

Metric	Test_value	Test_Bromley	Test_croydon
roc-auc	0.52	0.53	0.53
accuracy	0.59	0.6	0.61
precision	0.65	0.67	0.68
recall	0.76	0.785	0.77

The list of important features support our findings from the explortory analysis: 'total_land_area', 'ratio_proposed_land', 'max_height', 'vol_unit' were the only features that have the same directions of correlation with target variable. Probably, 'three_bed_units' got high 'gain' score because this feature differs significantly between the two boroughs, where the level of approval is also different

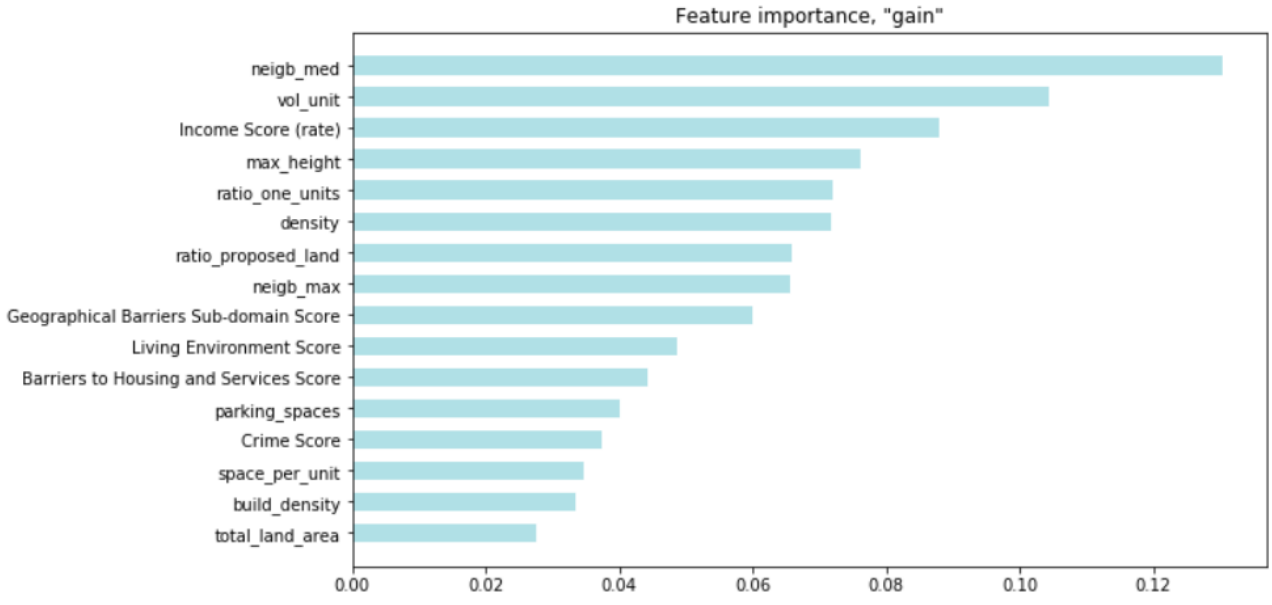


Figure 23

7.Discussion

7.1 Conclusion

This study aims to build a machine learning model to predict the probability of success of a planning application for a small block of flats in the London borough of Bromley. We used data of historical planning applications provided by Lumiere Property and the extra data collected from Croydon. Information on the built environment were added from Digimap and demographic data from Census 2011 to finalise our features. For methodology, we used QGIS and ArcGIS to code Croydon applications and finalise additional dataset. Three models were used for testing: Logistic regression, SVM, and XGboost. According to the results, we chose decision trees as the best model for Bromley data, which gave the 75% of accuracy of prediction with 77% on precision score. The maximum height of neighbourhoods (within 40 meters buffer), volume of unit and two IMD sub-domain data, geographical barriers score, and barriers to housing and services score are most important features in the final model.

7.2 Recommendations

From the undertaken research three conclusions can be made which hopefully will benefit the future work of Lumiere Property.

First of all, two of the three models Logistic Regression and Decision tree demonstrated a good level of precision (over 70%) on the test dataset. It means that only 30% percent of rejected applications were mistakenly named as successful. Hence, using any of these models will decrease the company's risks to fail by 70%.

Next, it is important to highlight that despite the good quality any of the built models cannot say the exact criteria of rejections in Bromley, as there are definitely other important factors beyond the ones used as features. It is possible that there is some historical relation between councils and developer and that some important neighbourhood characteristics have not been identified. However, this research allows us to identify trends and make recommendations related to both size of proposed building and locality. These can be taken into consideration for making application proposals in the future.

1. It is important to propose enough space for every planning unit because buildings with small units are mostly rejected
2. Following on the previous recommendation, this should not affect the size of building as buildings with larger area coverage than the average (>240 sq metres) are likely will not be approved
3. Buildings in neighbourhoods where the average height of property is bigger (over 9 meters on average) are more likely to be approved

4. Building in wealthier areas – where people can afford the more expensive housing – is less risky.

Finally, comparative analysis of Croydon and Bromley datasets signalled the difference existing in decision-making in two boroughs. Most likely councils assess applications by different criteria, therefore it is difficult to build one model for several boroughs in the future. Logistic Regression worked better on Bromley data alone. However, SVM showed small improvement of accuracy while using the two datasets together, but its results were still lower than ones of XGBoost model trained only on Bromley data. That is why, if the company decides to use ensemble tree modelling, we suggest using only Bromley data.

7.3 Limitations

There are several limitations to the project. The most significant one is the small number of observations available for the analysis. Since the applications in Bromley were limited to the ones received from Lumiere Property, it was decided to use Croydon data. Time was also a limiting factor of this project, as only limited time could be spared to data imputation and for testing the models. Due to the lack of time we decided to work with three models, however, this could be extended in the future.

7.4 Future work

Due to the lack of data points to use in modelling, this study only focussed on predicting the final status of the planning application. However, to improve the success rate of applications, the company needs to identify the reasons why a particular application got rejected. If sufficient data is available, a model could be developed which not only outputs the result of the application, but also the reason for rejection.

Currently, manual scraping was done to collect data for planning applications. Instead, a web scraper could be developed which would extract data from the planning applications portal using the application number. Natural Language Processing (NLP) could then be used on this data to identify the different attributes of the application.

8. References

Brown-Luthango, M., Makanga, P. and Smit, J. (2012). Towards Effective City Planning—The Case of Cape Town in Identifying Potential Housing Land. *Urban Forum*, 24(2), pp.189-203.

Carmon, N. (2001). Housing policy in Israel: Review, evaluation and lessons. *Israel Affairs*, 7(4), pp.181-208.

Chan, I. and Liu, A., 2018. Effects of neighborhood building density, height, greenspace, and cleanliness on indoor environment and health of building occupants. *Building and Environment*, 145, pp.213-222.

Eichholtz, P. and Lindenthal, T., 2014. Demographics, human capital, and the demand for housing. *Journal of Housing Economics*, 26, pp.19-32.

Géron (2019), *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly

Geyer, J., 2017. Housing demand and neighborhood choice with housing vouchers. *Journal of Urban Economics*, 99, pp.48-61.

Kleinhans, R. (2004). Social implications of housing diversification in urban renewal: A review of recent literature. *Journal of Housing and the Built Environment*, 19(4), pp.367-390.

Lindh, T. and Malmberg, B., 2006. Demography and housing demand—what can we learn from residential construction data?. *Journal of Population Economics*, 21(3), pp.521-539.

Porat, I. and Shach-Pinsly, D. (2019). Building morphometric analysis as a tool for urban renewal: Identifying post-Second World War mass public housing development potential. *Environment and Planning B: Urban Analytics and City Science*

Xgboost Developers. 2020. "Xgboost Release 1.1.0-SNAPSHOT Xgboost Developers."