

The social-housing allocation problem in Scotland

Design of a data-driven allocation model

Michael Redenti

Dissertation submitted in partial fulfilment for the
degree of Master of Science in Big Data

Computing Science and Mathematics

University of Stirling

Scotland

August 30, 2019

Abstract

The social-housing allocation problem in Scotland

Designing a data-driven allocation model

Michael Redenti

Abstract

In Scotland, the lack of a well-designed and effective social-housing allocation scheme is leading to substantial welfare losses: £19 million were lost through properties being empty in 2017/2018 [15]. We will analyse data from the Scottish Household Survey and investigate which socio-economic, cultural and demographic factors associate with residential satisfaction. Purposely, these results will aid in the design of two data-driven allocation models. One assigns housing units to applicants based on a similarity score, while the other is based on the prediction of a decision tree model that achieved 90% accuracy on test data. A system of distinct representatives can then be found by applying the Hungarian method [20].

Acknowledgements

I would like to acknowledge the contribution of my thesis supervisor, Dr. David Cairns of the University of Stirling, to the direction of the research. I would like to thank him for directing me on how to carry out research via the scientific method. His support, guidance and useful comments throughout this dissertation project are extremely appreciated.

Moreover, I would like to extend my thanks to all the Department of Computing Science and Mathematics for their role in coordinating and delivering this excellent Masters course.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this. I certify that this dissertation reports original work by me during my University project including/except for the following:

1. The Scottish Household Survey data [7] was published having gone through the cleaning stage already.
2. The code (feature selection algorithm) discussed in the methodology chapter was developed by me in Python through the aid of standard libraries (Pandas, Numpy, Scikit-Learn). However, the nature of the algorithm is a simple implementation of information theory equations.
3. One of the allocation models proposed in the Allocation chapter was inspired by many known content-based recommendation engines.

This project was carried out in conjunction with HomePointr CIC, a social enterprise devoted to the solve the issues of social-housing in Scotland. The nature of the problem was suggested by the CEO of this organisation: Fash Fasoro.

Signature: 

August 30, 2019

Contents

1	Introduction	1
2	Background and context	5
2.1	Brief history of social housing in Scotland	5
2.2	Social housing: application and allocation	6
2.3	The costs of a poorly designed allocation scheme	8
2.4	A new vision: HomePointr CIC	9
2.5	Related research	10
3	Methodology	15
3.1	The Scottish Household Survey _s	15
3.2	Data preparation	18
3.3	Data mining methods	21
4	Results	25
4.1	Model performance in time	25
4.2	Validating related research findings	29
5	Allocation	31
5.1	Similarity score - Content-based recommendation engine	32
5.2	Machine learning model	33
5.3	System of distinct representatives	34
6	Discussion	37

7 Conclusion

41

Chapter 1

Introduction

A welfare state's attempt to mitigate social inequalities is commonly achieved through the redistribution of wealth in the form of social goods and services. Unfortunately, these systems of provision often suffer from inefficiencies and, consequently, are inadequate to make this objective a fully functional reality.

On a global scale, “The market for public-housing exemplifies this phenomenon” [6]. Abusiveness, lack of maintenance, ineffective allocation policies, vacant properties, inadequate support and long bureaucracy time are some of the common problems associated with social-housing. The repercussions on the social-welfare of an economy are fairly substantial, not to mention the social burden on the community.

In this thesis, we will investigate the social-housing problem in Scotland, United Kingdom. Here, recent upward trends in the homeless population call for urgent action to be taken, both in the form of policies and investments [17].

Two key problems in the Scottish public-housing system are the lack of a fully centralised choice-based letting scheme and an inefficient allocation mechanism. These are believed to be the main reasons behind the worrying number of tenancy offers being rejected. While a choice-based letting scheme could perhaps improve the allocation system, this would arguably reduce the rate of rejections only partially. Indeed, in Scotland, although rejection rates for those housing associations that operate a choice-based letting scheme are lower than for housing providers with a

different application system, they are still present in a high number (more detail in the following sections). However, to answer this hypothesis with absolute certainty, it would be necessary to understand what drives these rejections explicitly. More precisely, it would be imperative to determine whether the problem can be attributed to a mismatch between the tenant's preferences and the dwelling itself or whether there are problems with that particular dwelling irrespective of the applicant's housing needs and aspirations. Unfortunately, data from the Scottish Housing Regulator [15] reports only aggregate statistics about the performance of social landlords. In particular, no information is reported about the reasons that drive individuals with housing needs to reject offered tenancies.

Hence, we turn to data from the Scottish Household Survey (SHS) [7]. The aim is to learn which factors drive individuals' residential satisfaction and understand the interaction among these variables. First and foremost, this would allow us to devise a parsimonious profile for applicants and properties, using only the most essential features. Then, the results of this investigation would allow us to build a model of allocation that maximises a predicted level of residential satisfaction, with the ultimate expectations of reducing the risk of sub-optimal allocations.

A great deal of research has been carried out to uncover the determinants of residential satisfaction among different socio-demographic groups. However, most of this research has been solely focused on specific time periods, on a yearly basis, without assessing whether socio-cultural and technological changes in time might affect residents' objective aspirations. To this end, this thesis will investigate all the SHSs available from 1999 until 2017. The argument is that, if no such changes have occurred, we will be able to potentially gain greater accuracy and statistical evidence during the model building phase.

Wealthy individuals dissatisfied with their living situations usually respond through migration in the pursue of a better alignment with their preferences and aspirations. However, this form of migration is likely to be unfeasible among individuals whom

are homeless and/or can not afford rental prices in the private market. Therefore, an efficient and effective public housing allocation system is of utmost importance.

Following this introduction, the thesis includes six chapters. The next chapter briefly discusses Scottish social-housing in time highlighting some of the controversial policies that still negatively impact the system today. The current bureaucratic journey that an applicant must go through when requesting a housing unit is outlined. The chapter concludes with an in-depth analysis of the related research into the topic of residential satisfaction and allocation models. Chapter 3 describes the data and the methodology adopted to carry out the investigation. In particular, emphasis is given to the analysis techniques used to validate the hypotheses. Chapter 4 reports the results of this study. Chapter 5 details the two proposed models of allocation and their evaluation on some test data. The thesis culminates with the discussion and conclusion chapters, where an evaluation of the investigation and suggestions for future work are given.

Chapter 2

Background and context

2.1 Brief history of social housing in Scotland

The first talks for the development of social housing in Britain date back to the 1850s when Prince Albert, Queen Victoria's husband, advanced the idea of affordable housing for the labouring classes in order to improve their living conditions. Later in 1919, the Government's first Housing Act was passed. This was set out to clear the city slums, invest in the first social-housing developments and improve the overall standard of living. Unfortunately, the Second World War put a lot of friction to this housing regeneration. Only after the war, the Scottish Government called for 50,000 homes per year to be built [11].

The 1980s mark what appeared to be an important turning point for individuals and families in socially rented accommodations: the UK Conservative government approved the Right To Buy (RTB) scheme, an empowering although controversial programme which facilitated the purchase of public homes with generous discounts [12]. While on the one hand it enabled housing ownership, on the other it was diminishing the available stock of social homes. This scheme would have a long-term severe impact. The Government could not match the more than one million loss of council homes with appropriate investments in the years to come. The effects are still being felt today [17].

Following many years of campaigning for better housing conditions and reforms, the

Right to Buy scheme finally came to an end on 1 August 2016.

In recent years, there has been an uprising trend in the demand for social housing as illustrated in Figure 2.1. In order to meet the future demand for social housing,

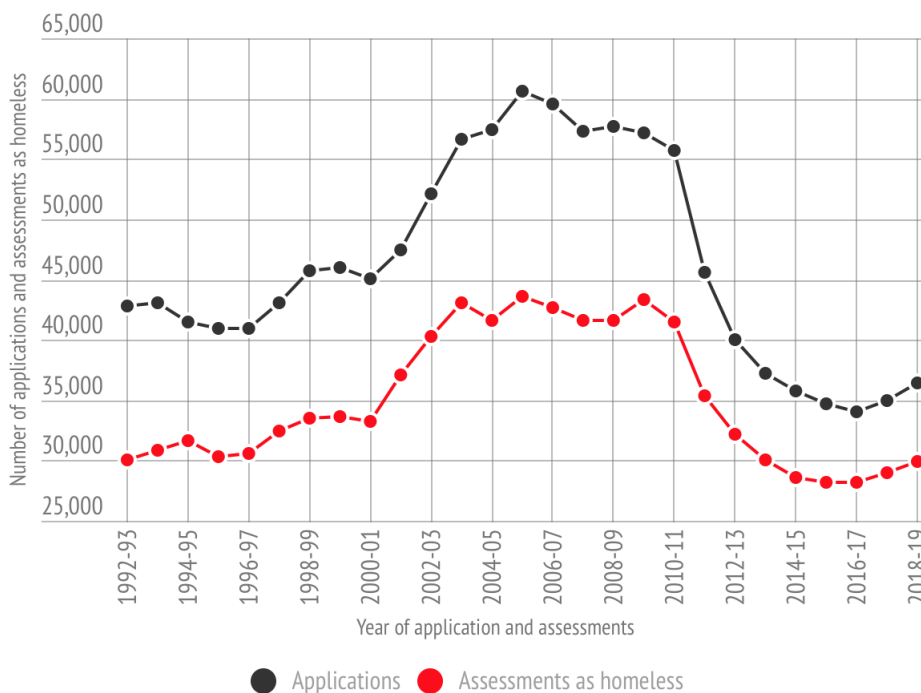


Figure 2.1: Caption - mention the picture is taken from Shelter Scotland

the Scottish government has invested £150 million over financial years 2018 till 2021 (the Building Scotland Fund).

2.2 Social housing: application and allocation

Applying for social housing

When applying for social housing, one must consider whether he/she is applying as homeless or not. If homeless, the individual or family must first of all make contact with the local council who oversees and validate the claim while temporary accommodation can be provided if necessary. Then, once the homelessness status is confirmed, the applicant joins a waiting list on a common housing register (council

level) for housing from the council, housing associations and cooperatives.

Other non-homeless individuals and families are still eligible for social housing. However, beside the option of joining a common housing register, they also have the possibility of applying directly to a housing association and express a preference for a particular property of interest. This is known as a choice-based letting scheme, typical of the private market rental process.

While there is a considerable advantage in choosing to join the waiting list for a particular property [6], the lack of a centralised system entails that a different application has to be made to each housing provider, a highly time-consuming process.

To streamline the bureaucratic journey of public-housing seekers, referral schemes have been put in place whereby voluntary organisations or agencies refer applicants to social landlords.

Allocation

The position of an applicant on the waiting list is determined by a points-based system which awards higher priority individuals or families groups where

- current living does not meet tolerable standards (ex: overcrowded);
- there are large families with children;
- there is a recognised homeless status/claim.

However, the allocation does not occur through a strict First-In-First-Out (FIFO) process. In fact, while reasonable preference is given to certain more vulnerable groups to ensure fairness, the Scottish law states that social landlords are responsible for developing their own final allocation policies and make decisions within these rules [6]. For instance, landlords could add other factors of their own such as health conditions or mobility impairments.

The law also sets out those factors social landlords must not take into account when allocating housing units. These are

- length of time you have lived in the current residency;
- any outstanding debt at a house where you were not the tenant;
- any rent or other money you owed to a landlord for a previous house which you have now repaid.

At the same time landlords must foresee that the property-applicant matching is optimal. As the following section demonstrates, this has not been an easy task.

2.3 The costs of a poorly designed allocation scheme

In this section we report the inefficiencies of the current housing allocation scheme through the analysis conducted on the biennial data from the Scottish Housing Regulator. The data contains aggregated statistics on the performance of all social landlords in Scotland.

In the years 2017/18 the average percentage of tenancy offers being rejected was around 42% for councils and 28% for other housing providers. This has been a common trend for the past years. It is worth pointing out once again that the higher rate of rejections for councils could be attributed to a lack of a choice-based letting scheme, option which is instead provided by housing associations and cooperatives. Figure 2.2 shows a breakdown of these figures for councils.

The chartered data does not report explicitly the reasons for these rejections nor applicants were asked to give motivation. Nonetheless, a few reasons can be conjectured. First and foremost, it is plausible that the effects of the RTB scheme are still being felt today. Precisely, it is likely that the more than one million housing units lost through the years had been the ones in the best location and overall conditions. Secondly, the lack of a centralised choice-based letting scheme for homeless people might lead to unfit or sub-optimal property-individual matching. This would likely cause properties being rejected once the viewing takes place. Inevitably, this has negative effects on the already long waiting lists (see Figure 2.3).

During these long waiting times, applicants' circumstances may change and as a

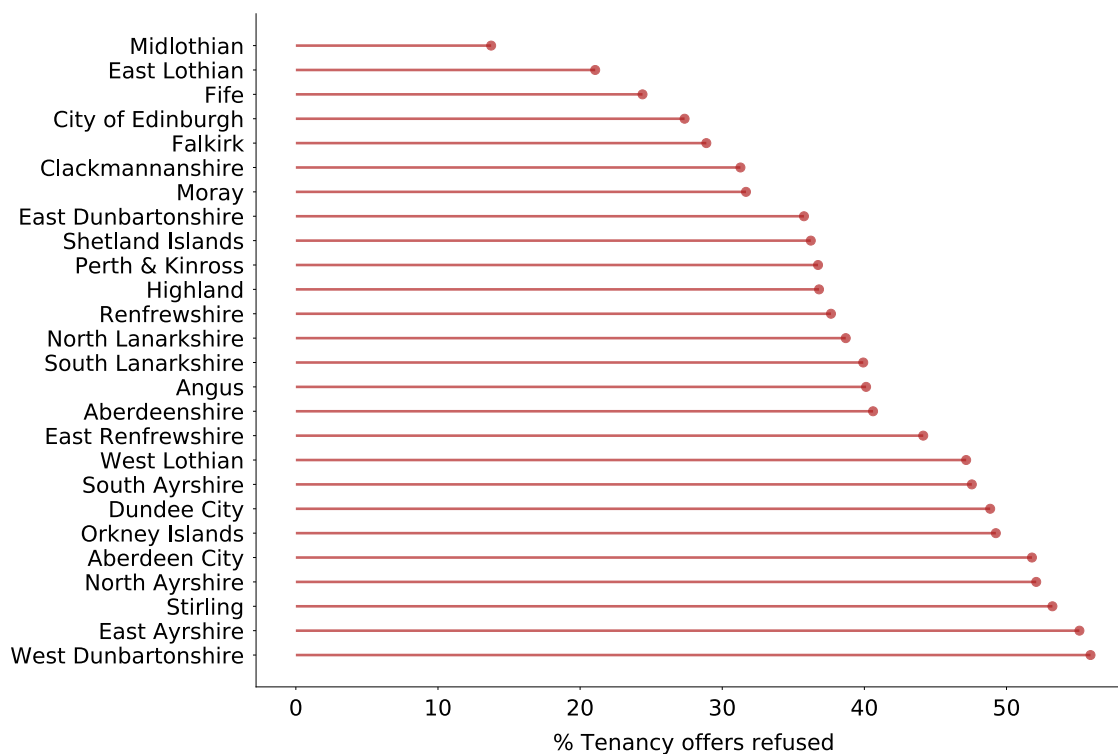


Figure 2.2: Evidence of the inefficiencies in the allocation system. The graph illustrates the high percentage of tenancy offers being rejected as a result of sub-optimal allocations.

result their preferences at the time of the application do not reflect their current status; once again assigned units are rejected. The importance of an allocation model that matches the suitability of an applicant to a certain housing unit by maximising a predicted level of satisfaction will be the focus of this thesis. According to the Scottish Housing Regulator, the rent lost through properties being empty in 2016/2017 was £90 million.

2.4 A new vision: HomePointr CIC

HomePointr CIC is a social enterprise established to design an online platform to facilitate the social-housing allocation process. It aims to bridge the gap between referral agencies and housing providers. Ultimately, the aim of this project is to

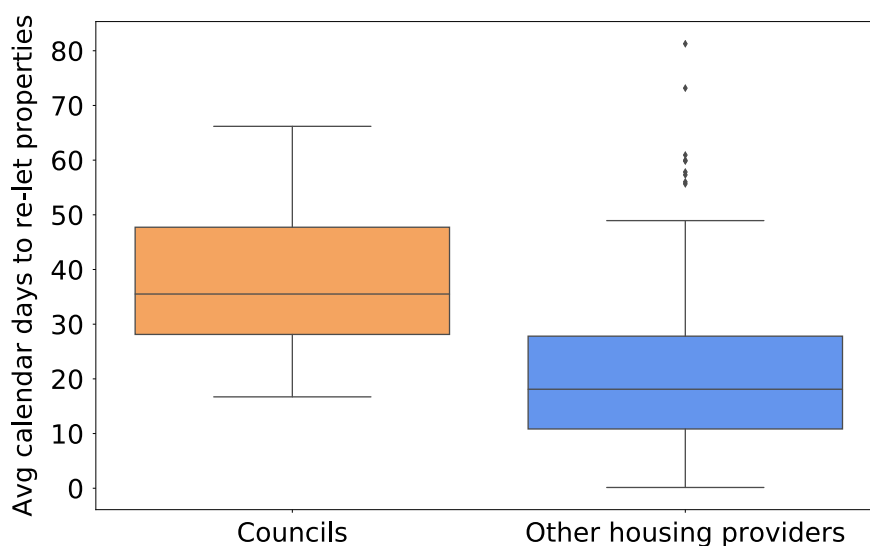


Figure 2.3: The average calendar days to re-let properties is higher for councils than for other housing providers, who instead operate a partial choice-based letting scheme.

integrate such platform with one of the proposed automated allocation systems (see Chapter 5) hoping to achieve optimal allocations and reduce the currently high rate of tenancy rejections.

A centralised choice-based letting system

The online platform would combine the two current types of application processes used by social landlords in Scotland into a single centralised choice-based letting system. In a nutshell, the idea is to bring the efficient design of the private sector rental marketplace such as AirBNB to public-housing, while ensuring a non-strict priority based letting scheme as envisioned by the Scottish law.

2.5 Related research

Residential satisfaction has been a long standing area of focus in academic research of various disciplines, being a very complex theme affected by a variety of socio-

demographic, behavioural and cultural factors [10]. Scientist of all disciplines have come to an agreement upon its interpretation as one's subjective assessment of the difference between his/her actual and aspired-to living situation [9].

Policy makers benefit enormously from research studies into this topic as they provide directions for the design of new housing policies while serving as a tool for monitoring the progress of existing ones. This section briefly discusses the major research results which will simultaneously motivate the importance of our approach to solve the social-housing misallocation problem.

According to Schneider [16] , when it comes to one's assessment of quality of life measures there is no consistent relationships between objective descriptions and subjective perceptions. For example, residents of a neighbourhood with a poor reputation can still be satisfied with their housing arrangements. Permentier et al. analyse household survey data from the city of Utrecht, Netherlands, and find differences between the determinants of neighbourhood satisfaction and of the perceived reputation of that same neighbourhood [14]. Perhaps, it could be reasoned that when respondents are asked about how others see their neighbourhood they make this evaluation based on objective measures. Consequently, they are likely to have a different and more rational perspective.

The subjective nature of one's assessment of household satisfaction is a key point that has motivated the use of survey data as the preferred method of analysis for these type of studies rather than objective measures [10, 14, 3].

The association of certain indicators with residential satisfaction is more obvious than others. For instance, people with higher income and/or owning a property with greater market value are likely to be more satisfied with their living arrangements [10].

With regards to socio-demographic variables, older people tend to be more satisfied with their tenancy than younger people [3]. Amerigo and Aragonés claim that such may be an indication of individuals' acceptance of their housing status over time [1].

Households in poor living conditions manifest a similar behaviour, they adjust their level of satisfaction in time as a response to financial limitations and consequent lack of choice [14, 1].

Empirical studies have identified a number of other important determinants such as available space in the house, ethnic composition, neighbourhood and rural/urban status. The strong association with neighbourhood satisfaction is reported in numerous studies [22, 16, 14]. Fried and Gleicher and Western et al. argue that one's own living situation assessment is likely to include its immediate surroundings [4, 21]. However, which specific aspects of neighbourhood are taken into account by individuals or households are still very much unclear [14].

The literature also presents some conflicting results; specific variables may have significant effects on the response in some studies but not in others or the direction of the effects may be opposite. For example, while some authors have provided evidence that no effect can be attributed to genders, others have reported that single men are less likely to be satisfied than women. Also, Kasarda and Janowitz have provided evidence that duration of residence has a positive effect on neighbourhood satisfaction [8]. On the other hand Onibokun, in his study of subsidised housing in Canada, concludes that longer stays were associated with lower levels of neighbourhood satisfaction [13]. However, such a conflict can be easily resolved by noting the social status of the different samples. The sample in Onibokun's analysis are social households who, as a result of a First-In-First-Out allocation scheme are less likely to be allocated housing units that meet their needs and aspirations [18]. Given that financial circumstances might be the limiting factor in their relocation to better housing conditions, the aspiration for better housing might culminate in lost hope and adaptation in the long term.

In [10], Max Lu argues that these and other inconsistencies in the literature can be partly attributed to the different samples being analysed and partly to the erroneous methodologies of analysis. Regression techniques that do not take into account the ordinal nature of the response variable are inappropriate. While we accept this hy-

pothesis, we also advance the possibility that there is a time component when it comes to socio-cultural predictors whose interaction may change in determining the levels of residential satisfaction. In this case, surveys across different years would inevitably lead to different results not only because of the different samples, survey designs and methods of analysis. To this regards we will pursue, in this report, such investigation by analysing survey data spanning a period of six years.

Since most of the research has been focused on residential satisfaction in the private market, there is little mention on whether there are any differences with predictors of residential satisfaction in social-housing. Nonetheless, Permentier et al. report that people who experience freedom in the choice of their property and neighbourhood are going to be more satisfied [14]. Indeed this lack of choice is believed to explain partly why social tenants are generally less satisfied with their dwelling than homeowners of the private market. Moreover, this finding motivates the mission of HomePointr CIC in delivering a centralised choice-based letting system in the social-housing market, which coupled with an effective data-driven allocation scheme should mitigate the current inefficiencies.

Allocation

Neil Thakral discusses extensively the drawbacks of a First-In-First-Out (FIFO) non-choice based allocation model and how it leads to sub-optimal allocations [6]. The author proposes a mechanism where applicants choose among a set of waiting lists, each representing a different housing unit and estimated waiting time. This allows applicants to make a free, informed decision and an offer prior to the unit becoming available.

Chapter 3

Methodology

The first part of this chapter will provide a description of the design, collection and composition of the Scottish Household Survey (SHS) data [7] across the years. Then, information about the data processing and analysis techniques used to carry out the investigation into the determinants of residential satisfaction follows.

3.1 The Scottish Household Survey_s

The SHS is a continuous survey of a sample of households across Scotland. It provides a wide range of detailed information on the composition, characteristics, attitudes and behaviour of residents in both private (70%) and social housing (30%). The interview takes place face-to-face interview with the Highest Income Householder (HIH) or their partner. Interviewers are equipped with a Computer Assisted Personal Interviewing (CAPI) which is used to collect the answers to the questionnaire.

The survey began in 1999 and up until 2011 followed a biennial and fairly consistent design. However, from 2012 onwards (data has been published up until 2017), the survey was substantially redesigned to include elements of the Scottish House Condition Survey and new subjective social indicators [5]. Indeed, only approximately 20 questions about typical demographic information such as gender, age, income and social status have been asked in a consistent manner across all these surveys

(see Table 3.1).

Survey year	Survey year												
	2017	2016	2015	2014	2013	2012	2011	2009/10	2007/08	2005/06	2003/04	2001/02	1999/00
2017	1	78.17	96.36	69.31	74.29	65.68	14.72	14.72	14.03	9.62	6.85	6.52	6.03
2016	79.33	1	77.14	86.5	70.6	82.6	26.33	26.29	24.43	18.59	11.96	10.97	10.76
2015	95.58	75.39	1	72.88	77.41	68.35	17.11	17.03	15.41	10.23	7.4	6.5	6
2014	68.59	84.33	72.7	1	81.26	95.15	29.99	29.87	27.41	19.05	12.51	11.02	10.57
2013	81.99	76.81	86.13	90.63	1	89.28	22.96	22.82	20.30	13.68	10.58	9.4	8.28
2012	64.94	80.47	68.13	95.07	79.99	31.5	31.34	28.64	19.72	12.94	11.29	10.60	
2011	18.19	32.05	21.32	37.44	25.70	39.36	1	99.29	83.71	54.28	39.46	34.87	30.79
2009/10	17.91	31.51	20.89	36.72	25.16	38.56	97.76	1	83.97	53.94	39.20	34.49	30.22
2007/08	13.00	22.30	14.40	25.67	17.05	26.84	62.79	63.96	1	59.50	44.57	38.03	27.03
2005/06	9.6	18.28	10.30	19.22	12.38	19.91	43.86	44.27	64.11	1	74.78	54.74	35.64
2003/04	6.56	11.28	7.14	12.10	9.17	12.53	30.58	30.85	46.05	71.71	1.0	76.87	51.56
2001/02	7.68	12.73	7.83	13.11	10.04	13.45	33.23	33.39	48.34	64.58	94.56	1	64.29
1999/00	9.14	16.05	9.26	16.18	11.36	16.24	37.73	37.61	44.16	54.04	81.53	82.64	1

Table 3.1: Percentage of common questions among surveys. Note that each survey has a different number of features, hence the asymmetry.

In furtherance of obtaining a representative information of households living all over the country, the sample covers all 32 local authorities in Scotland. The surveys' sample size has been chosen by taking into account Scotland's population size at that time. The number of households interviewed varies from a minimum of 10,304 to a maximum of 10,658 households across these surveys. Consequently, this provides a reliable nation wide representation of households' living conditions across Scotland. A non-irrelevant portion of questions might only be asked of a third of the households in the sample and/or on a biennial basis. This explains the high degree of missing data as illustrated in Figure 3.1.

As previously mentioned, each survey consists of wide-ranging topics from household composition, economical and social status, attitudes, daily activities and behaviour. Table 3.2 illustrates a general overview the typical survey topics from 2012 onward.

Table 3.2: Typical topics and example questions covered in the 2012-2017 SHS

Household composition and characteristics				
Type	Single adult	Single parent	Large Family	Single pensioner
Classification	Urban	Rural		
Sex (HIH)	Male	Female		
Ethnicity (HIH)	White	Minority ethnic group	Refused	
Religion (HIH)	Church of Scotland	Roman Catholic	Other Christian	Other
Housing				
Tenure	Owned outright	Rent	Buying with mortgage help	
Landlord type	Council	Housing association	Private landlord	Other
Neighbourhood and communities				
Neighbourhood rating	Very good	Fairly good	Fairly poor	Very poor
Sense of community	Very strongly	Fairly strongly	Fairly poor	Very poor
Sound pollution	Very common	Fairly common	Not very common	Not at all common
Neighbourhood disputes	Very common	Fairly common	Not very common	Not at all common
Economic activity				
Status (HIH)	Employed full time	Employed part time	Self employed	Unemployed and seeking work
Highest qualification	University degree	'O'-Grade	A level	Higher National Diploma

	Local Services			
Satisfaction	Satisfied	Dissatisfied	Neither	No opinion
Local bus usage	Every day	2/3 timed per week	Once a week	Once a month
Health service	Very satisfied	Fairly satisfied	Fairly dissatisfied	Very dissatisfied
Police service	Very satisfied	Fairly satisfied	Fairly dissatisfied	Very dissatisfied

Anonymisation

The survey data was provided subject to non-disclosure with non-registered third parties. Use of these data and related publishing is allowed for academic purposes, but would require a special permission for its use in a commercial setting.

Accordingly, access to the data has been password protected in order to elude the risk of misappropriation. Further, the data had already been anonymised to preserve the identity of the respondents. Households are uniquely identified by a randomly generated ID. There is no personal or other information that would allow to directly trace back the identity or location of the household.

The data consist of a tab-separated file and a supplementary HTML file containing a translation of the encoded column names and their values.

3.2 Data preparation

The survey question that we will use as a measure of residential satisfaction is “On the whole, how satisfied or dissatisfied are you with this house/flat?”, encoded as ‘pa1’ in the file. The response to this question takes one of six possible answers:

- very satisfied;
- fairly satisfied;

- neither satisfied nor dissatisfied;
- fairly dissatisfied;
- very dissatisfied;
- no opinion.

The cleaning process had already been taken care off by the publishers of the survey [7].

Missing data

Given that only the SHSs from 2012 until 2017 include the question of interest, we will only consider these data sets in our analysis. Across these surveys, the response rate to this question is consistent, ranging around 30% (see Table 3.3).

	Survey Year					
	2012	2013	2014	2015	2016	2017
Sample size	10642	10628	10622	10304	10454	10658
% Responses	31.86	32.37	32.35	31.95	32.13	30.12
% Responses without 'no opinion'	31.84	32.36	32.32	31.93	32.13	30.10

Table 3.3: Percentage of households who responded to the household satisfaction question.

Now, the relevance of a 'no opinion' response is equivalent to no response at all in terms of our investigation. Moreover, the percentage of such responses is fairly irrelevant, less than 0.03% across all surveys. Consequently, any such observations have been discarded from further analysis.

Partly as a result of the design of this survey (see Section 2.1), there is a high degree of missing data for the remaining features too. Figure 3.1 shows the distribution of missing values across the survey questions for each year. The distributions are fairly consistent across years with only a small proportion of features having zero

missing values, but with a high proportion, approximately 1000 questions, having no response at all. In our investigation we will only consider those feature whose

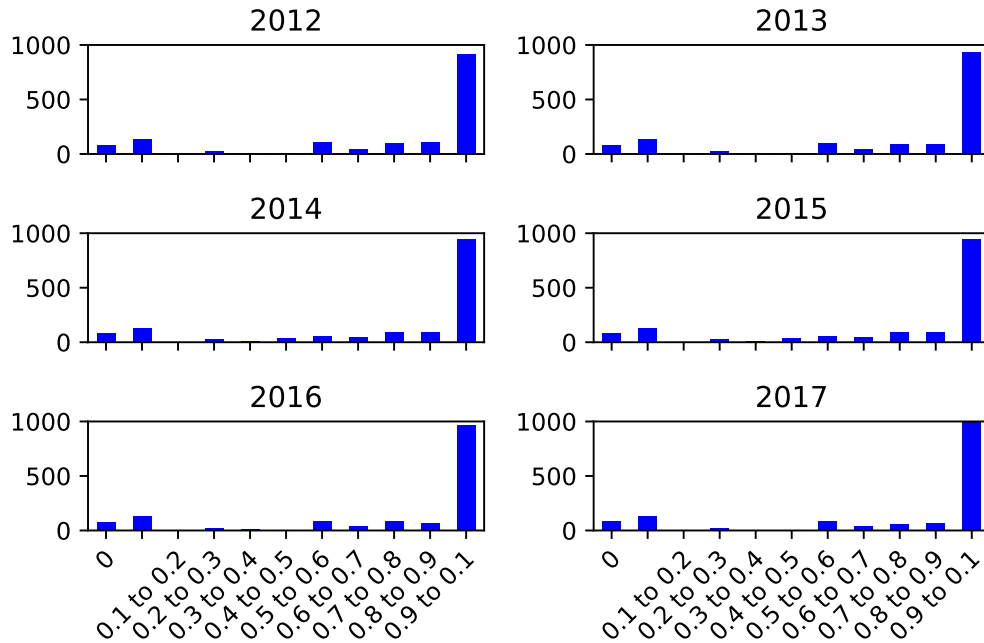


Figure 3.1: Number missing values per column feature. 1505 total common features.

percentage of missing values is below a sensible threshold of 30%. (Stratified random sampling or SMOTE will be used to carry out the estimation of these missing values. - perhaps run the analysis with different thresholds)

Data transformation

The data is mostly nominal in nature with no relevant features having a continuous domain. For ease with the model building stage and subsequent interpretation, variables such as age and income were categorised. Participants were grouped into young (16 - 35), middle-aged (36 - 55) and old (> 56).

Irrelevant information such as year of the survey (it this was used as a feature to observe whether it is a factor, we might miss its relevance if using forward feature selection), household ID and other were removed from the model training data.

In our analysis we will also investigate how the modelling results change as we convert the response into binary classes, that is ‘Satisfied’ (very satisfied, fairly

satisfied) and ‘Not Satisfied’. From a logical perspective the ‘Not satisfied nor dissatisfied’ response is converted appropriately. Moreover, in terms of an allocation model, this would lead to a lower risk of sub-optimal allocations.

3.3 Data mining methods

Clearly, a greater level of statistical evidence and a more reliable accuracy score of a supervised model would be achieved if we were to merge all the surveys into one data set. However, we need to be cautious and take into account the potential presence of a time dependency factor that could otherwise cloud the results of hypothesis tests and/or affect the accuracy of a model trained in this manner. Indeed, it is plausible that, for instance, as a result of socio-cultural dynamics and technological innovations, one’s assessment of residential satisfaction today might take into account different or more factors than one’s assessment compared to previous years. Moreover, beside the interest in this particular phenomenon, we would demand that a model trained on all the surveys would perform at least as well as a model trained on the single and, perhaps, most recent survey, more likely to approximate the current underlying model.

This line of argument is examined as follows. Firstly, we will extract from each survey the greatest determinants (features) of residential satisfaction and compare them. Subsequently, we will train a model (decision tree or else) based on these highest predictors for each survey year. We will inspect whether the interaction among their values remains unchanged and whether a consistent level of accuracy is achieved on the other data sets. This will provide us with enough evidence to determine the validity or not of such hypothesis.

Dealing with unbalanced data

The data is highly unbalanced with approximately 90% of Scottish households being satisfied with their living arrangements. The binary conversion of the response can not mitigate this unbalancedness (see Figure 3.2). So, in order to avoid model bias

in favour of the overly represented class (or classes), we will under sample each class according to the number of observations in the minority class.

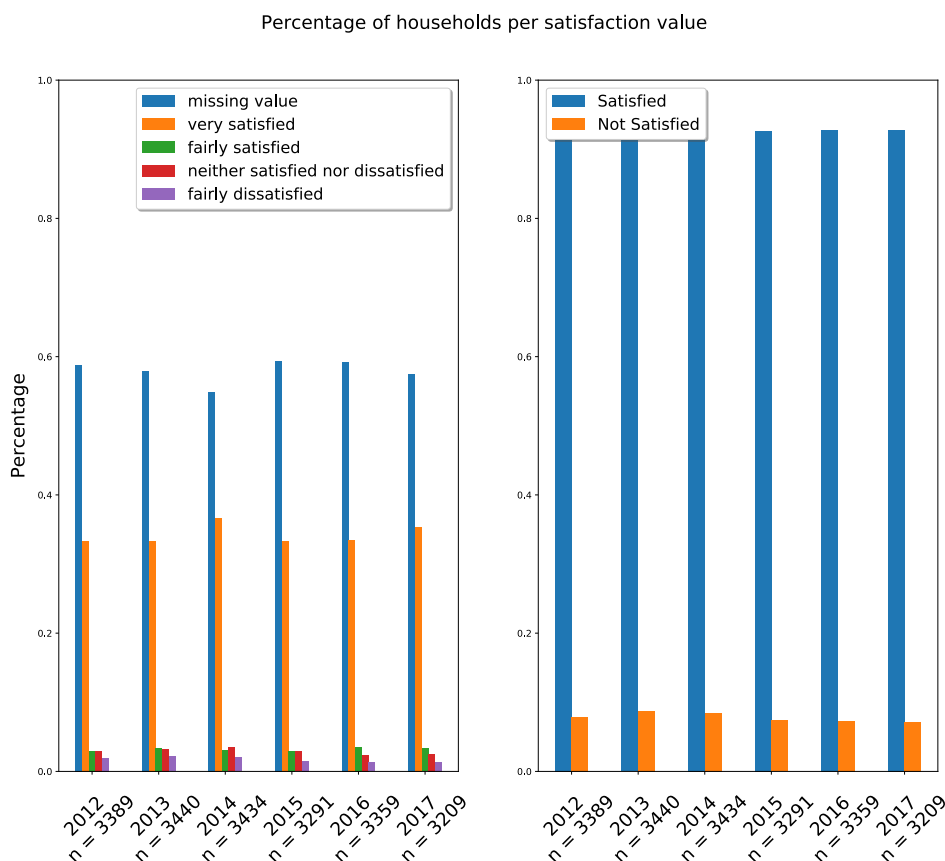


Figure 3.2: Non-binary response

Feature Forward Selection

Given the high dimensionality of the data, we will extract the top predictors of residential satisfaction via a feature forward selection algorithm. Conditional entropy will be used the selection criteria.

In information theory, conditional entropy H quantifies the uncertainty in the outcome of a variable Y given information about a single or multiple variables X . The equation (discrete features and response) for conditional entropy is as follows:

$$H(X|Y) = - \sum_{x \in X, y \in Y} p(x, y) \log_b \frac{p(x, y)}{p(x)},$$

where the base of the logarithm b is chosen to be the same as the number of classes so that this quantity ranges between 0 and 1.

A tolerance of 0.01 in the decline of the conditional entropy is chosen as the stopping criteria or else when the information score reaches zero.

Now, since features that are highly correlated with any of the currently selected features by the algorithm will not aid any significant contribution to the information score, these will likely be discarded during the selection process. Consequently, in order to extract all highest predictors, we will run the algorithm multiple times, eliminating the features from the previous run at each stage.

Chapter 4

Results

4.1 Model performance in time

The predictive power of the greatest predictors of residential satisfaction in any three year period remains reasonably stable in time.

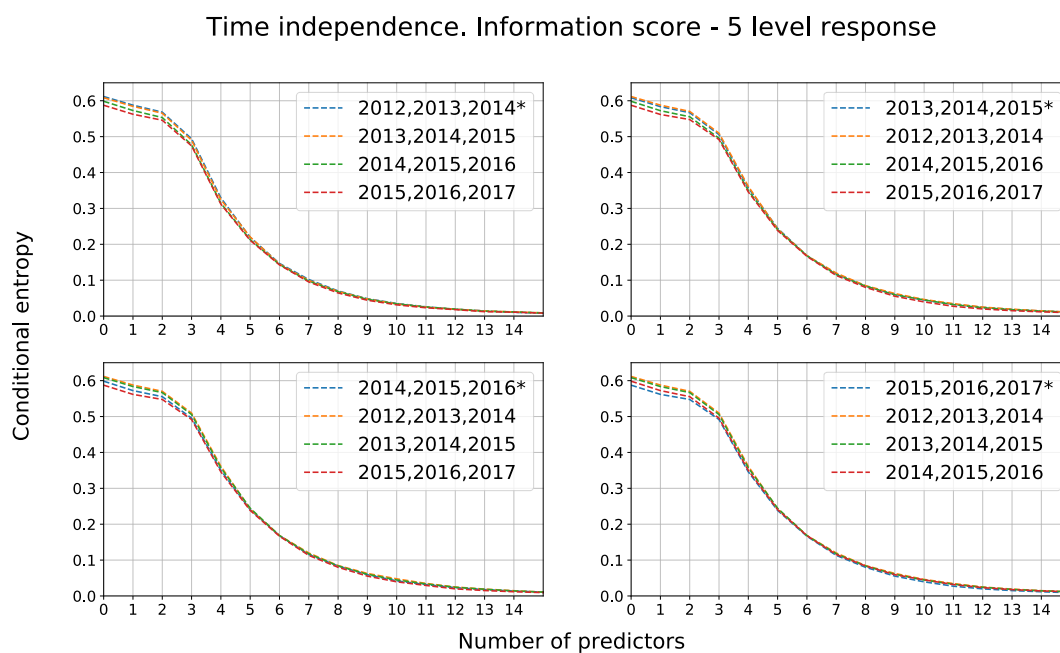


Figure 4.1

Each graph in Figure 4.1 shows the conditional entropy score as a function of those

predictors that have been forward selected on the indicated (*) three years survey period.

Given the differences in the samples, under a time-invariant assumption we would expect the same rate of change for each curve. Moreover, any relatively small differences in the rate of change can be attributed to noise in the observations.

This same phenomenon is observed also when transforming the response into binary (see Figure 4.2).

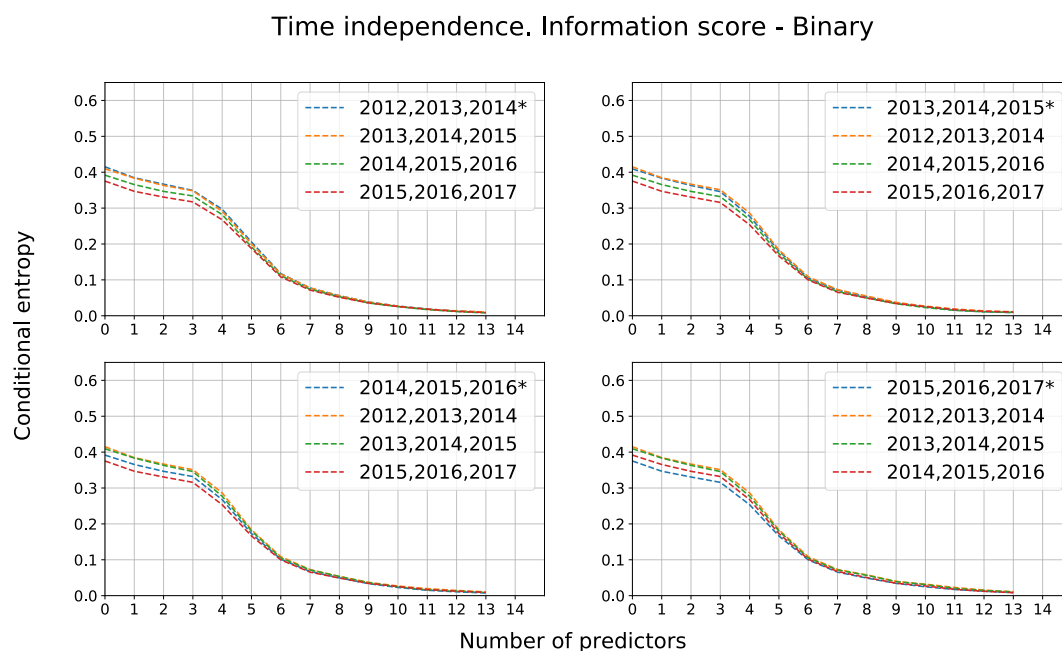


Figure 4.2: Caption

The accuracy of a decision tree model also does not decline neither on future nor on past observations, as illustrated in Figure 4.3. This would be expected, given there is a certain correspondence between a conditional entropy score of a set of features and the accuracy of a machine learning model.

A binary decision tree was trained for each of the four consecutive three years survey periods from 2012 to 2017. The test data (20%) was also under sampled to match the same proportion of observations across the other years survey period. As a result, comparisons among different models' scores are justified. In Figure 4.3, the accuracy of the trained model on its own remaining test data is shown in the legend

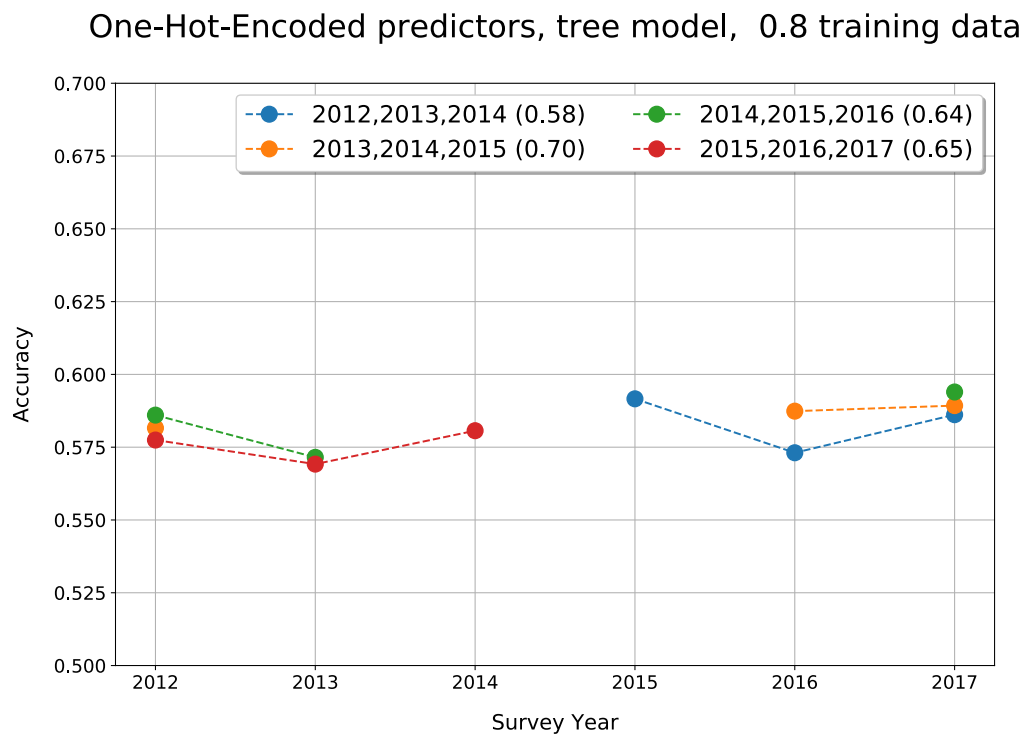


Figure 4.3: Caption:

in brackets.

Clearly, if this time invariant property is going to be persist also outwith the period considered, in a real setting, there would be no significant improvement by re-training the model including new survey data nor its performance would be altered.

Time consistency in the determinants of residential satisfaction

So far we have simply observed that the accuracy of a model trained on any of the available consecutive three year period data does not lose its predictive power in time. However, we are still left to determine whether the predictive features of each model are different among each other, that is whether different time periods are associated with different predictors. This would then clarify whether the slightly better accuracy score of certain models are a result of different predictors or, simply, noise.

Each graph in Figure 4.4 illustrates once again the conditional entropy score for

5 level response (left) and binary (right)

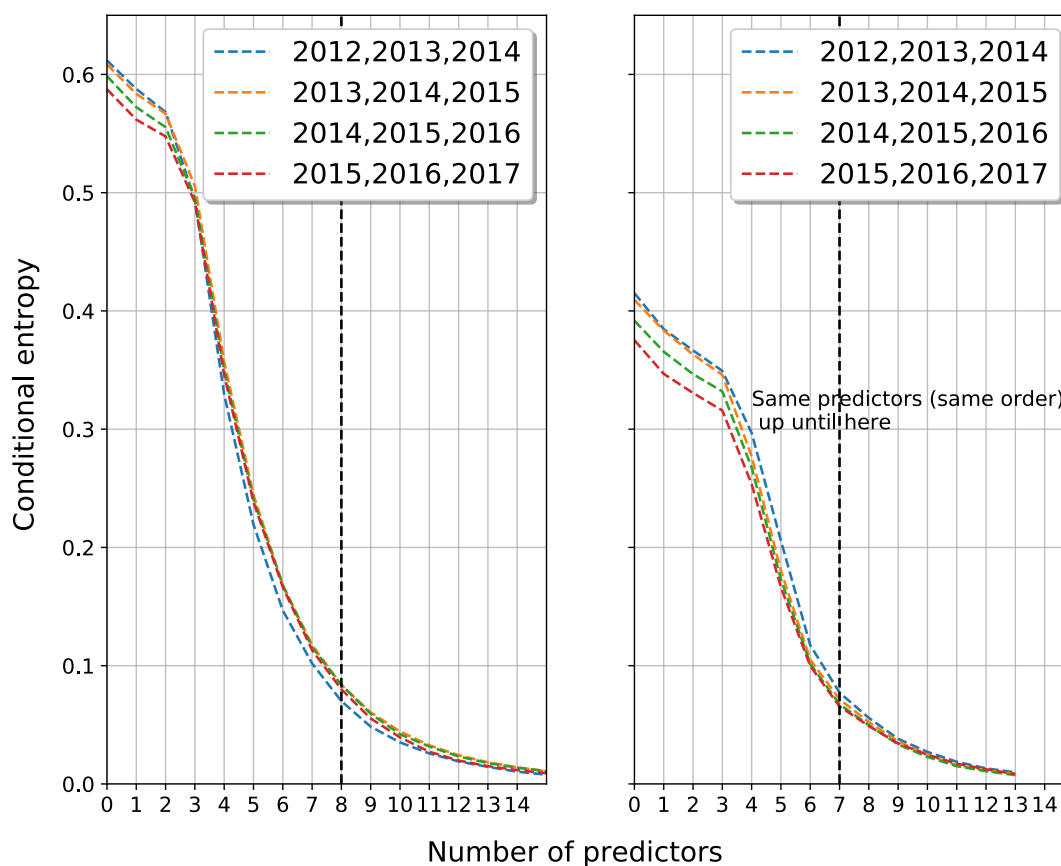


Figure 4.4

each three year period, but now the feature forward selection has been carried out separately for each data set. The black vertical dashed line indicates when at least one change in the predictors and/or their order has occurred. This happens for low conditional entropy values, that is when a model is likely to start over-fitting the training data. However, while the forward selected features are the same across the 3-year periods, we also need to verify that the interaction among their values also remain invariant. To this end, we compare the four different tree models whose accuracy scores were reported in Figure 4.3. Indeed, the underlying predictors of residential satisfaction and their interaction are consistent.

4.2 Validating related research findings

In this section we will investigate the findings reported in the related research, Section 2.1. Moreover, given that most of the investigations consulted for this thesis date back to at least 2010 or over, this validation will give us an intuition of whether predictors of residential satisfaction have changed over a greater time period than the one available from our survey data.

Hypothesis testing on all the available survey data (2012-2017) will be performed to investigate the statistical association between predictive features and the response of interest. The appropriate statistical test for nominal data is χ^2 test.

		Residential Satisfaction				
		Satisfied	Not satisfied	df	χ^2	p-value
Neighbourhood subjective rating	Good	16481 (93.72%)	1103	1	1044.87	< 0.001**
	Poor	645 (65.34%)	342			
Community belonging	Strongly	13826 (94.67%)	778	1	596.15	< 0.001 **
	Not strongly	3203 (82.80 %)	665			
Overall size of the house/flat	Satisfied	17325 (95.16%)	881	1	2350.70	< 0.001 **
	Not satisfied	1214 (63.79%)	689			
Opinion about number of rooms	Too few	1691 (78.35%)	467	2	667.27	< 0.001 **
	Too many	1496 (90.55%)	156			
	About right	15361 (94.16%)	951			

Table 4.1: Validating related research findings.

Table 4.1 reports the contingency tables, with the χ^2 and p-value scores, among various subjective and objective indicators, partly suggested by other research into the topic, with response of interest.

Neighbourhood satisfaction and feeling of community belonging are significantly associated with residential satisfaction at the 5% and 1% level. Also, residents who tend to be satisfied with the size of the household are likely to be more satisfied with their living arrangements. This association is still present even when we transform the “Overall satisfaction with the size of the house/flat” into an objective and quantitative measure: $\frac{\text{number of bedrooms}}{\text{family size}}$.

Chapter 5


Allocation


Ultimately, the aim of this investigation is to devise a new data-driven and more effective social-housing allocation mechanism to be integrated into a web application (HomePointr CIC - see the introduction for more details). The app would support three types of users: referral agencies, individuals and landlords.

In a nutshell, the typical scenario would involve a landlord receiving x different applications across y of its listed properties. Then the allocation model would automatically find the combination, matching housing units to applicants, that maximises a predicted score/level of residential satisfaction.


Matching applicants to housing units

First and foremost, determining which features are most predictive of residential satisfaction will allow us to draft a parsimonious profile for both applicants and properties. Figure 5.1 illustrates what a potential user profile would look like, similarly for a property profile.





Dann Jaskolski Edit

 **Housing Preferences**

RENT MIN	RENT MAX	NUMBER OF BEDROOMS
£ 450	£ 750	2
LOCATION	FAMILY SIZE	
Falkirk	3	

Figure 5.1: An applicant's preferences contains features of various nature.

In the following two sections, two potential allocation models are proposed and described in detail.

5.1 Similarity score - Content-based recommendation engine

The first model is very much inspired by the typical content-based recommendation engine. From [2] we observe that the general model for a similarity based recom-

mendation engine can be represented concisely as follows:

$$Sim(a, p) = \sum_i^n sim(a_i, p_i) * w_i, \quad (5.1)$$

that is a weighted sum of similarity scores between corresponding property-applicant features.

With regards to numeric features calculating a similarity score based on a distance function is quite straightforward. However, when considering categorical feature, a form of encoding is needed. The easiest approach is to adopt a matching score (1 or 0) but there is not really a concept of distance with regards to the other values. For example, when considering a post-code we should rank the non-matching post-codes in terms of their closeness to the desired value by the user. (easier for binary variables)

$$sim(u_i, p_i) = Manhattan, EuclideanDistance \quad (5.2)$$

$$sim(u_i, p_i) = \begin{cases} 1, & \text{if } u_i = p_i \\ 0, & \text{otherwise} \end{cases} \quad (5.3)$$

The weights could be the relative frequency of the filtering. However, the user does not access the platform and so this is not feasible. The weights would determine how much the user values each property.

The above approach is good for categorical features with low cardinality that do not have an inherent hierarchy.

For numeric features such as money for example, in this case we have to do some processing. For example we could give a 1 to properties that are within 10% example of the price suggesting a threshold above which the applicant would not consider renting anyhow.

5.2 Machine learning model

A second model we propose is a decision tree model trained on all the available survey data, 2012 to 2017. As previously discussed, feature transformation has been carried out, where possible, in order to derive objective metrics from social indicators. Then,

this accounts for the different circumstances between the respondents of the survey and prospective social housing applicants.

A split ratio of 80:20 has been chosen for the training and test data respectively, and a maximum depth of 8 has been set during cross-validation, starting from a depth of 4. The model reaches an accuracy of 60.54% and 94% for the five level and binary response respectively.

5.3 System of distinct representatives

In essence, any one of the two models proposed positions the matching of a property to an applicant on a scale between 0 and 1, if we consider the binary response for the machine learning model.

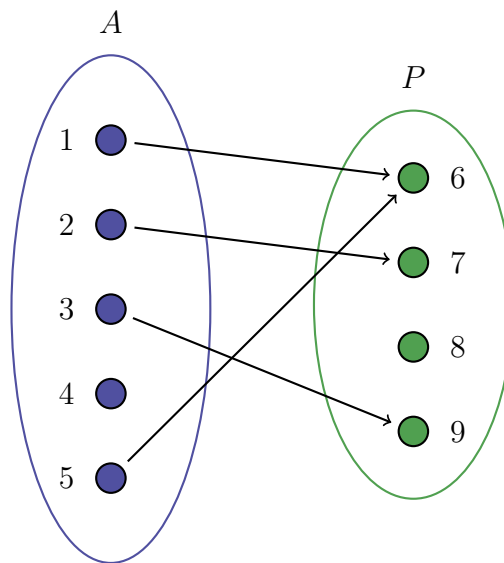


Figure 5.2: A set of applicants expressing preferences for housing from a social landlord.

However, we still need to address if and how a social landlord should allocate properties to applicants based on these predictive scores. For simplicity, let us consider the following scenario: a social landlord would like to allocate $|P|$ properties among a set of applicants A . Then, the aim of an allocation model would be to find the matching between properties and applicants such that the sum of each predicted

level of residential satisfaction is maximised.

This can be more easily understood and represented as a directed bipartite graph, as illustrated in Figure 5.2. An algorithm (Hungarian method) would be implemented so that a maximum matching is found, that is a set of edges in the corresponding bipartite graph that share no endpoints and it is the one of maximum size [19].

Chapter 6

Discussion

When compared to other investigations into the determinants of residential satisfaction, our analysis considered survey data which spanned multiple years (2012-2017). This allowed us to research whether the factors that households consider when expressing their dwelling satisfaction level would change in time, as a result of socio-cultural and/or technological changes for example. Although we did not find any significant changes, the time period considered might be sufficient to draw any conclusions nor can we infer with confidence this would not happen over even longer time periods. In other words, such changes might occur over greater life cycles.

Nevertheless, given the consistency of these predictors over the time period considered, we were able to make use of all the survey data available when examining relationships between subjective and objective indicators with the response of interest.

With regards to the determinants of residential satisfaction, our results relate to the expectations of this study based on the major findings in the literature.

The strong statistical association with neighbourhood satisfaction ($\chi^2 = 1044.87$, $df = 1$, $p\text{-value} < 0.01$) is a well known and established result. Max Lu argues that when respondents are asked about the level of satisfaction with their dwelling, they are likely to consider its immediate surroundings [9].

Despite, according to the literature, residential satisfaction being a highly subjective

matter that often does not find agreement with objective measures, we demonstrate that it is still possible to build a predictive model of a satisfactory accuracy (better than chance), based solely on objective metrics. Indeed, through appropriate transformation of subjective social indicators, we can engineer measures that can partly explain the response of interest. For instance, consider the strong relationship between the respondents' level of satisfaction with the size of the dwelling and the response of interest. Since such a question could not logically be asked to prospective social-housing applicants, we devised a new quantitative and objective metric related to it: $\frac{\text{number of bedrooms in the property}}{\text{household size}}$. The association was still strong following this conversion.

In Chapter 4 we outlined two models of residential satisfaction, both characterised by the ability of generating a score between an housing unit and an applicant.

The first model we proposed was simply a weighted sum of similarity scores between corresponding applicant and property features. The features to be included as part of the property/applicant profile were extracted using a feature forwards selection algorithm. Conditional entropy was used as the selection criteria. One of the advantages of using conditional entropy as the selection criteria selection is that highly correlated variables with the response and with the predictors that had been already selected, is that they will likely be discarded. This is extremely useful in order to achieve a parsimonious set of features. Moreover, in terms of profile building for the online platform, this would entail less information would be needed to be supplied by users whilst achieving still great results.

With regards to the second model, a machine learning model, we had to validate a few hypothesis prior to using all the available survey data for training. The investigation can really be broken down into two aims, in order of importance. First of all, we required that a model's performance based on the top predictors in any year does not decrease over subsequent years nor the interactions among their values change. The interactions were the same.

While the purpose of this model is to serve as an automated allocation tool, we realise its potential use also as a recommendation engine. Indeed, recommendation systems play a vital role in improving the user experience on e-commerce platforms and consequently increase their retention. It is effective in the way it reduces information overload by filtering the items of interest which are believed to be most relevant to the current user. Achieving a satisfactory and functioning level of personalisation requires engineering ways to monitor the user online interaction. For example, many recommendation engine start by saving the most recently viewed items and then make new suggestions based on a similarity score between those recently viewed properties and non-viewed properties.

Future work should focus in the implementation of this allocation scheme and monitor its performance. It would be wise to trial it with one of the social landlords and gather evidence of its performance.

Chapter 7

Conclusion

The social housing misallocation problem, the consequent tenancy offers rejected and the long waiting lists have inevitable repercussions on the economy as a whole. Improving the allocation mechanism will have a positive effect in the redistribution of wealth and allow to achieve greater social cohesion.

The Scottish Housing Regulator survey should not limit themselves to report only the aggregate performance of social landlords across Scotland. They should go more in depth with regards to why tenancy offers are being rejected. Tenants should be individually asked why the tenancy was rejected. This would have led to a better understanding and evidence of the causes. Consequently, our suggestion that a more effective allocation model will lead to fewer tenancies being rejected still needs to be fully validated.

The integration of one of the data-driven allocation system into a choice-based letting scheme online platform for social-housing might overcome some of the great inefficiencies related to sub-optimal allocation of tenancies. Moreover, given the high volume of properties social landlords might have at their disposal, automating this process will be extremely beneficial in terms of management simplification. And speed up the re-letting times.

The well-being of underprivileged people improves through a smooth access to property rights, a decrease in the time spent in temporary accommodation and a quicker access to better living conditions.

Bibliography

- [1] Maria Amerigo and Juan Aragones. “Residential satisfaction in council housing”. In: *Journal of Environmental Psychology* 10 (Dec. 1990), pp. 313–325. DOI: [10.1016/S0272-4944\(05\)80031-3](https://doi.org/10.1016/S0272-4944(05)80031-3).
- [2] Keith Bradley, Rachael Rafter, and Barry Smyth. “Case-Based User Profiling for Content Personalisation”. In: vol. 1892. June 2000. DOI: [10.1007/3-540-44595-1_7](https://doi.org/10.1007/3-540-44595-1_7).
- [3] Earle. Davis and Margret Fine-Davis. “Predictors of satisfaction with housing and neighbourhood: A nationwide study in the Republic of Ireland”. In: *Social Indicators Research* 9.4 (Dec. 1981), pp. 477–494. ISSN: 1573-0921. DOI: [10.1007/BF00286349](https://doi.org/10.1007/BF00286349). URL: <https://doi.org/10.1007/BF00286349>.
- [4] Marc Fried and Peggy Gleicher. “Some Sources of Residential Satisfaction in an Urban Slum”. In: *Journal of the American Institute of Planners* 27.4 (1961), pp. 305–315. DOI: [10.1080/01944366108978363](https://doi.org/10.1080/01944366108978363). eprint: <https://doi.org/10.1080/01944366108978363>. URL: <https://doi.org/10.1080/01944366108978363>.
- [5] Scottish Government. *Scottish household survey 2015: annual report*. 2016. URL: <https://www.gov.scot/publications/scotlands-people-results-2015-scottish-household-survey/pages/2/>.
- [6] Scottish Government. *Social housing: Housing management*. 2017. URL: <https://www.gov.scot/policies/social-housing/housing-management/>.

-
- [7] Scottish Government Ipsos MORI. “Scottish Household Survey, 1999 - 2017.” In: *[data collection]. 2nd Edition. UK Data Service. SN: 8333* (). eprint: <http://doi.org/10.5255/UKDA-SN-8333-2>. URL: <http://doi.org/10.5255/UKDA-SN-8333-2>.
- [8] John D. Kasarda and Morris Janowitz. “Community Attachment in Mass Society”. In: *American Sociological Review* 39.3 (1974), pp. 328–339. ISSN: 00031224. URL: <http://www.jstor.org/stable/2094293>.
- [9] M Lu. “Analyzing Migration Decisionmaking: Relationships between Residential Satisfaction, Mobility Intentions, and Moving Behavior”. In: *Environment and Planning A: Economy and Space* 30.8 (1998), pp. 1473–1495. DOI: [10.1068/a301473](https://doi.org/10.1068/a301473). eprint: <https://doi.org/10.1068/a301473>. URL: <https://doi.org/10.1068/a301473>.
- [10] Max Lu. “Determinants of Residential Satisfaction: Ordered Logit vs. Regression Models”. In: *Growth and Change* 30.2 (Mar. 1999), pp. 264–287. ISSN: 0017-4815. DOI: [10.1111/0017-4815.00113](https://doi.org/10.1111/0017-4815.00113). URL: <https://doi.org/10.1111/0017-4815.00113>.
- [11] Martin Stilwell MA. *Social Housing History*. 2015. URL: http://www.scotlandhousingcrisis.org.uk/scotlands_housing_crisis/.
- [12] Kim McKee. “The End of the Right to Buy and the Future of Social Housing in Scotland”. In: *Local Economy* 25.4 (2010), pp. 319–327. DOI: [10.1080/02690942.2010.498956](https://doi.org/10.1080/02690942.2010.498956). eprint: <https://doi.org/10.1080/02690942.2010.498956>. URL: <https://doi.org/10.1080/02690942.2010.498956>.
- [13] Adepoju G. Onibokun. “Social System Correlates of Residential Satisfaction”. In: *Environment and Behavior* 8.3 (1976), pp. 323–344. DOI: [10.1177/136327527600800301](https://doi.org/10.1177/136327527600800301). eprint: <https://doi.org/10.1177/136327527600800301>. URL: <https://doi.org/10.1177/136327527600800301>.
- [14] Matthieu Permentier, Gideon Bolt, and Maarten van Ham. “Determinants of Neighbourhood Satisfaction and Perception of Neighbourhood Reputation”. In: *Urban Studies* 48.5 (2011), pp. 977–996. DOI: [10.1177/0042098010367860](https://doi.org/10.1177/0042098010367860).

- eprint: <https://doi.org/10.1177/0042098010367860>. URL: <https://doi.org/10.1177/0042098010367860>.
- [15] Scottish Housing Regulator. *AFS data – all social landlords dataset, 2017/2018*. 2019. URL: <https://www.scottishhousingregulator.gov.uk/find-and-compare-%20landlords/statistical-information..>
- [16] Mark Schneider. “The ”Quality of Life” and Social Indicators Research”. In: *Public Administration Review* 36.3 (1976), pp. 297–305. ISSN: 00333352, 15406210. URL: <http://www.jstor.org/stable/974587>.
- [17] Shelter Scotland. *Scotland’s Housing Crisis*. 2018. URL: http://www.scotlandhousingcrisis.org.uk/scotlands_housing_crisis/.
- [18] Neil Thakral. “The Public-Housing Allocation Problem : Theory and Evidence from Pittsburgh”. In: 2016.
- [19] Dartmouth University. *Notes*. 2018. URL: <https://www.cs.dartmouth.edu/~ac/Teach/CS105-Winter05/Notes/kavathekar-scribe.pdf>.
- [20] Harvard University. *Notes*. 2018. URL: http://www.math.harvard.edu/archive/20_spring_05/handouts/assignment_overheads.pdf.
- [21] John S. Western, Peter D. Weldon, and Tan Tsu Haung. “Housing and Satisfaction With Environment in Singapore”. In: *Journal of the American Institute of Planners* 40.3 (1974), pp. 201–208. DOI: [10.1080/01944367408977469](https://doi.org/10.1080/01944367408977469). eprint: <https://doi.org/10.1080/01944367408977469>. URL: <https://doi.org/10.1080/01944367408977469>.
- [22] Julian Wolpert. “Migration as an adjustment to environmental stress”. In: *Journal of social issues* 22.4 (1966), pp. 92–102.