National College of Ireland

Novel Genetic Algorithms for Optimization of House Price Prediction: USA

MSc Research Project
Data Analytics

Ian Patterson
Student ID: 18124917

School of Computing
National College of Ireland

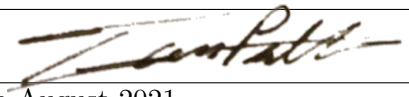
Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ian Patterson
Student ID:	18124917
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/08/2021
Project Title:	Novel Genetic Algorithms for Optimization of House Price Prediction: USA
Word Count:	9246
Page Count:	32

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Novel Genetic Algorithms for Optimization of House Price Prediction: USA

Ian Patterson
18124917

Abstract

Genetic Algorithms and their variants are popular methods of feature selection optimisation, with applications ranging from medical image denoising to structural engineering and stock market prediction. Innovations in the Genetic Algorithm itself have been limited, and not biologically-inspired. In this investigation, standard and novel biologically-inspired Genetic Algorithms are used to optimize the feature selection component of a machine learning-based House Price Prediction application. This analysis found the novel Co-Location and Multi-Chromosomal Genetic Algorithms to achieve superior optimization of XGBoost, Linear Regression and Decision Tree-mediated prediction. The Co-location Genetic Algorithm reduced Decision Tree prediction error by an additional 20%, as compared to the Standard Genetic Algorithm. The Multi-Chromosomal Genetic Algorithm reduced the required number of features for XGBoost and Decision Tree to achieve optimal prediction error, by 31% and 32% respectively, as compared to the Standard Genetic Algorithm. These optimal performances were also achieved with fewer generations of evolution required, corresponding to reduced computational cost. This finding has implications for all house price prediction analyses in which feature selection is mediated by genetic algorithms. Further investigation of the utility of these novel genetic algorithms across different domains may have implications for all applications of genetic algorithms, regardless of industry or application.

1 Introduction

Genetic Algorithms (GA) belong to a family of optimisation algorithms based on the principles of evolution and natural selection. They are effective in a range of applications but excel in the optimisation of np-hard problems, such as the Travelling Salesman Problem. Machine Learning-related applications include hyperparameter tuning and feature selection. The optimisation of feature selection is especially important, as machine learning problems grow more complex and challenging. Minimising the feature set size reduces computational costs and risk of bias during modelling (Bzdok et al.; 2018). The housing market in the United States of America constitutes a total value of 32.6 trillion dollars, according to Zillow, an online property marketplace (Zillow_Research; 2020). 2020 saw the largest annual growth in history, at 2.2 trillion, corresponding to an increase in property sales of 5.9% as compared to 2019 (Zillow_Research; 2020). Further, Zillow estimates that 2021 property sales will increase by a further 21.9%. These statistics highlights the value in property investing, and continued growth of the market. With growth in the market, the potential for profit increases, as does the value to be gained by understanding

and predicting market behaviour. GA have successfully improved the ability of machine learning models to predict property market changes (Dong et al.; 2020) and improve accuracy of property price predictions (Liu and Liu; 2019; Su et al.; 2021).

The motivation for this investigation is two-tiered. First, benchmarking of existing and novel GA for optimised feature selection may yield improved house price prediction models. This can allow for developers to better understand and therefore build properties with attributes most valued by customers, improving customer satisfaction and profits. Improved predictive models would also confer an advantage in the purchase and reselling of properties. Secondly, the novel GA protocols designed and investigated here may improve upon the state-of-the-art GA. A new optimisation technique would potentially improve upon any and all existing applications of Standard Genetic Algorithm (SGA), across research domains.

1.1 Research Question, Objectives and Contributions

The lack of biologically inspired enhancements to the GA, and the ubiquitous and important application of the optimisation method across industries and domains, motivates improvement of the methodology. This project therefore aims to answer the research question:
To what extent do additional biologically-inspired enhancements improve industry standard Genetic Algorithmic efficiency, in a United States property price prediction application.

The total objectives of this analysis are summarised in Table 1. Objectives 1, 2 and 3 aim to establish a context for the house price prediction domain; these include a literature review, sourcing and processing of an appropriate analysis dataset and the exploratory analysis of this dataset. The outputs of these objectives will be the basis from which the subsequent investigations are completed. Objectives 4 to 8 (inclusive) comprise the data mining, machine learning, optimisation, and implementation of novel genetic algorithms, for house price prediction. Due to the complexity and novelty of the genetic algorithms formulated during this study, these objectives (6, 7) include sub-objectives and sub-investigations.

The major contributions of this investigation are the novel GA methodologies designed and explored here, and the identification of the machine learning models that most accurately predict house price error. The novel GA methodologies may be used by researchers from disparate domains to improve the efficiency of their feature selection optimisation. The house price prediction error model may contribute to more efficient and profitable price prediction model for real estate agents.

The insights garnered from the exploratory analysis of house price prediction error in Los Angeles comprise a minor contribution of this study. This exploratory analysis will allow Zillow to better understand the market in which they operate, and where their price prediction algorithm is least effective.

This study is limited in that GA performance is only measured against a single dataset. It may be that the GA explored here performed especially well, but only for this dataset. Also, the data included here is restricted to recent years in the Los Angeles US housing market. It is unclear whether the insights garnered through this analysis are

generalisable to other housing markets. Also, as this investigation is aimed at optimizing the Zillow Zestimate house pricing model, the insights from the machine learning components may be less useful for non-Zillow pricing. Although GA may perform well in optimising the architecture of neural networks, there is insufficient time to include neural networks in this study. Therefore, it may be possible to improve predictor accuracy and efficiency through GA-optimisation of neural networks.

Table 1 - Research Objectives

Obj.	Objective Description	Techniques	Evaluation
1	Review of the Genetic Algorithms and House Price Prediction Error		
2	Pre-Processing, Feature Engineering and Missing Value Investigation	Cleaning Imputation	Visualisations
3	Exploratory Analysis of price prediction errors in Los Angeles	Hypothesis tests	Statistics Visualisation
4	Unoptimized House Price Prediction, and benchmarking of unoptimized modelling techniques.	Random Forest XGBoost Linear Regression Decision Tree	Mean Absolute Error
5	Implementation of Standard GA for Feature Selection, using superior models from objective 4	Genetic Algorithm	Mean Absolute Error Features Used
6	**Design and Implementation of a novel Gene Co-location GA for Feature Selection**		
6a	Correlation analysis of features and Construction of a Network	Correlation Graph Theory	Spearman Rho
6b	Implementation of a Travelling Salesman -Type Genetic Algorithm for optimisation of network graph	Genetic Algorithm	Total Distance
6c	Implementation of gene co-location algorithm with superior predictive models from objective 4	Genetic Algorithm	Mean Absolute Error Features Used
7	**Design and Implementation of a novel multi-chromosomal GA for Feature Selection**		
7a	Translate optimal solution from obj. 6b into a 2D Matrix and identify groups of related genes via K means Clustering.	KMeans Clustering	Total Sum Squared Error
7b	Implementation of Multi-Chromosomal GA for Feature Selection, using superior model from objective 4	Genetic Algorithm	Mean Absolute Error
8	Compare and benchmark GA performance among novel variants explored here	Visualisations	Descriptive Statistics

The structure of this technical report is as follows: A review of the related academic work will be presented in Section 2. This will be followed by an overview of the Methodology used in this investigation, in Section 3. The Design Specification will be presented in Section 4. This will be followed by the Implementation, Results and Evaluation of each investigation, in Section 5. This section will also include a Discussion of the results. Section 6 will include the Conclusions of the investigation will be presented, alongside Limitations and Future Work.

2 Related Work

This review of the related work in the literature includes an investigation into and critique of feature selection techniques used for optimisation of machine learning and house price prediction analyses. As this study focuses on improving Genetic Algorithm performance, this family of algorithms is also investigated. The review concludes with an examination of the house price prediction domain, and a critique of machine learning and optimisation techniques that have found success. The scope of this review is restricted to developments in the feature selection, genetic algorithm and house price prediction domains from the years 2013 to 2021.

2.1 Feature Selection and Techniques

Feature selection is a technique used to select the best features to be used during the modelling and machine learning steps of an analysis. The aim is to reduce the number of features incorporated into the model such that the most predictive features remain, and the least predictive (including redundant and counter-productive) features are removed. This is especially relevant today, as datasets become larger and more heterogeneous; there is more irrelevant and erroneous data obscuring the meaningful data (Li et al.; 2017). Reducing the number of features included in a model leads to simpler models that are more interpretable (Blanquero et al.; 2019). It also reduces the amount of noise, improves the computation speed during modelling and reduces risk of overfitting and biasing models (Muñoz-Romero et al.; 2020). Feature selection is key for producing predictive or classification models with reproducible and more generalisable performance. This is vital in modern fields of biotechnology such as microarray screening of genetic disease, where the vast majority of genes do not contribute to disease (Saeid et al.; 2020). This field involves the analysis of thousands of genes in the human genome (each gene is a feature), feature selection is required to "ignore" irrelevant features that interfere with important features (Hambali et al.; 2020).

2.1.1 Critique of Feature Selection Techniques

Feature selection techniques can be objectively grouped into three different types: filter methods, wrapper methods and embedded/hybrid methods. Filter methods are based on the individual characteristic features. These methods are simpler and less computationally expensive, but they do not account for interactions and relationships between component features and the learning approach. Filter methods rely on basic statistics describing the properties of the feature set, such as correlation and variance, and their more generic and quick application comes at a cost of decreased accuracy (Jha and Saha; 2021).

Wrapper methods incorporate machine learning or evolutionary computation into the selection process. This allows for detection of relationships between variables, and therefore are more likely to achieve greater accuracy, at the cost of computational expense. This was demonstrated when filter and wrapper feature selection methods were applied to a Parkinson's disease classification problem. The wrapper method outperformed the filter method by 21% in classification accuracy (wrapper: 88%; filter: 67%) (Gündüz; 2019). Although it is expected that filter methods are always less computationally expensive, it has been demonstrated that this may be rectified through choice of machine learning model. In a comparison of filter and wrapper methods aimed at optimising classification

across multiple datasets (ranging from beans to flowers to lung disease), Wrappers using Decision Tree (DT) and Naïve Bayes performed faster than the filter methods. Although, these were the exception, as the filter methods were faster than K Nearest Neighbours and Support Vector Machines (Xue et al.; 2015).

The third approach, Embedded methods, are a hybrid of the two; specific learning algorithms and heuristics are applied to speed up computation while preserving most of the Wrapper method's superior performance. An embedded method was used to optimise the feature set used for understanding and predicting crashes, automatically generated through an IT system. This method first involved a shortlisting of features using an information gain-ratio filter method, and further delineated the optimal features using a weighted balanced distribution adaptation wrapper method. This was embedded method was found to outperform 25 filter methods across 7 different experiments (Xu et al.; 2021).

2.2 Genetic Algorithms and Recent Advances

GA are composed of 5 main steps: population initialisation, selection, fitness evaluation, crossover, and mutation. These steps can be altered and reconfigured such that the benefits of GA are enhanced and the pitfalls, such as pre-mature convergence, are avoided. Aside from configuration of the stages of GA, heuristics can be used to incorporate expert and industry knowledge into the optimisation process. This allows for a non-naïve and more efficient search of the solution space. A convolutional neural network was constructed for medical image denoising. Initialising the population of the hyperparameter tuning GA with parameters that had been successful in the literature, such as the ADAM optimiser, accelerated optimisation and reduced incidence of pre-mature convergence (Liu and Liu; 2019). Similarly, incorporation of train scheduling best practices into the population initialisation of a scheduling optimisation problem improved total efficiency, as measured by time required to complete the schedule (Vlašić et al.; 2019). Recent advances in GA architecture have included the dynamic variation of crossover and mutation rates. Increasing the crossover and mutation rates when diversity tends to stagnate is a means by which additional genetic variability can be introduced into the population. While this may increase risk of losing the best performing solutions, it reduces risk of pre-mature convergence. Dynamically varying the mutation rate substantially reduced the number of generations required to optimise the travelling salesman problem (Xu et al.; 2018), while dynamically altering crossover rates was similarly successful (Hussain et al.; 2017). Interestingly, varying crossover rates across populations, in a multi-population ga improved total diversity and resulted in improved algorithmic efficiency. This indicates that accelerating the key genetic exchange process of GA improves algorithmic efficiency (Wang et al.; 2018).

2.3 Review of the United States Housing Market and House Price Prediction

In the United States, 5.7 million homes were sold via Zillow in the year 2020, constituting a 5.9% growth year-on-year. This growth is projected to increase to 22% (7 million homes) in 2021 (Zillow_Research; 2020). As the value and number of homes being sold increases year on year, the value in predicting property prices consequently increases. Understanding the factors that drive prices is key to leveraging this market. Given that price reflects what people value, the phenomenon is difficult and complex to model. The

current industry practice is the hedonic model, in which prices are determined by the structural characteristics of and amenities local to the property (Owusu-Ansah; 2013). However, this view is limited, and does not make use of more complex and nuanced data available to real estate agents and would-be modellers. Data describing social environment factors such as human mobility patterns, and aesthetic factors such as the local landscape and property architecture may dictate price (Du et al.; 2018; Kang et al.; 2020).

2.3.1 Critique of Techniques Used for House Price Prediction

Given the value and profit derived from an accurate understanding of the real estate market and property valuation, researchers have tackled the problem for decades. Relevance Vector Machines, Linear Regression (LR), Bayesian Linear Regression and Gaussian Probabilistic Regression were applied to a London property price dataset. Of these statistical techniques, Gaussian Regression was superior. Success was attributed to the flexibility and non-linearity of the model (Ng; 2015). Accuracies superior to those achieved by statistical methods in the London sample above, were achieved by tree-based methods in a Chinese sample. DT, Random Forest (RF) and Gradient Boosted Trees (GBT) were compared, and it was found that the RF performed best (Truong et al.; 2020). However, RF was outperformed by a stacked ensemble of RF and GBT. However, as these machine learning techniques were used on a separate sample, they may not be directly comparable to the London study above. More modern machine learning techniques such as Neural Networks have potential to vastly improve upon the accuracies achieved by simpler methods, due to their increased complexity and ability to integrate novel data types. A deep neural network that incorporated street view images of the properties in question, and thereby introducing aesthetic to the pricing model, achieved supreme accuracy RMSE < 0.1 in a US Dataset (Kang et al.; 2020). However, this dataset consisted of 60,000 properties in a relatively restricted location. The computational expense may prohibit such a protocol for larger datasets comprised of millions of properties across disparate geographies.

GA have also been used to optimise house price prediction, to achieve accuracies greater than what would be achieved through complex machine learning models alone. A GA optimised a Log-periodic power model that predicted turning point in house price trends in a Chinese sample (Dong et al.; 2020). The efficient performance of the model was attributed to the evolution of solutions across multiple populations. A GA was also used to optimise a Long Short-Term Memory (LSTM) neural network, that achieved a Mean Absolute Percentage Error of 6% in a Unites States sample of x properties. The error achieved by LSTM alone was twice this error (12%) (Liu and Liu; 2019). Finally, in a study of automated property appraisals, a GA performed two key functions. It first acted to optimise the features selected for the optimal model performance, reducing the feature set from 65 to 8 key features. It also performed a multi-objective optimisation of the diversity and accuracy achieved by the Gradient Boosted Regression model, used to predict house prices. This optimised model achieved an R^2 of 0.9, compared to the unoptimized model's top R^2 of 0.87. Although this difference is marginal, it was achieved using fewer features, and is therefore more generalisable and less computationally expensive (Su et al.; 2021).

2.4 Summary of Findings, Identified Gaps and Conclusion

This literature review found that the most effective machine learning algorithms for the prediction of house prices, across multiple geographies, were deep neural networks and other ensemble classifiers such as RF and GBT. The predictive power of these models were improved through optimisation of feature selection. The findings of this literature review are summarised in Table 2. However, no study of house price prediction was identified to have used either the Multi-Chromosomal Genetic Algorithm (MCGA) or Gene Co-Location Genetic Algorithm (CLGA) methodologies. Moreover, when researching the state-of-the-art and novel GA methodologies across domains, still no study was found to use these techniques. Therefore, the design and implementation of these novel methodologies may rectify a significant gap in the literature, with the potential for applications across industries and research domains. Additionally, this literature review did not identify any house price prediction studies using the Zillow Zestimate dataset, used in this analysis. Therefore, this investigation may also be the first of its kind in that regard.

This literature review comprised completion of Objective 1 (See Section 1.1, Table 1).

Table 2 - Comparison of Techniques Use

Citation	Location	Sample	Period	Techniques	Optimisation
(Ng, 2015)	London	2.4 million	1995-2013	Relevance Vector Machines, Linear Regression, Bayesian Linear Regression, Gaussian Probabilistic Regression	None
(Truong et al., 2020)	China	300,000	2009-2018	Decision Tree, Random Forest, Gradient Boosted Trees, Stacked Ensemble (Random Forest + Gradient Boosted Trees)	None
(Kang et al., 2020)	USA	22,000	2014-2019	Deep Neural Network, Support Vector Machine, Long Short-Term Memory	None
(Dong et al., 2020)	China	30,000	2017	Log Periodic Power Model	Genetic Algorithm
(Liu & Liu, 2019)	China	664	2011-2017	Long Short-Term Memory, Support Vector Machine, Artificial Neural Network	Genetic Algorithm
(Su et al., 2021)	China	19,000	2000-2017	Genetic Algorithm + Gradient Boosted Tree, Gradient Boosted Tree	Genetic Algorithm
This Analysis	USA	2 million	2017-2018	Gradient Boosted Tree, Random Forest, Linear Regression, Decision Tree	Multiple Novel Genetic Algorithms

3 Multi-Chromosomal and Co-Location Genetic Algorithm Methodology

The methodology used for this investigation is termed the Multi-Chromosomal and Co-Location Genetic Algorithm Methodology. This has been adapted from the popular Knowledge Discovery in Databases methodology, to meet the specific objectives of the analysis. In this section, the preparatory methodological steps will be described briefly. The Implementation will include descriptions of the steps that follow the pre-processing and Transformation stage, in greater detail.

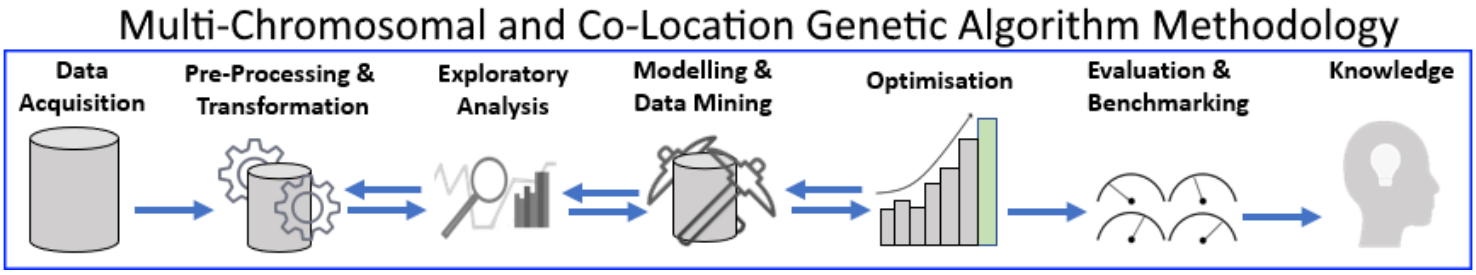


Figure 1: *Summary of the Multi-Chromosomal and Co-location Genetic Algorithm Methodology, adapted from the Knowledge Discovery in Databases methodology.*

3.1 Data Acquisition

The data used in this analysis was sourced from a Zillow database, published to Kaggle as part of a competition (Kaggle; 2018). This data was composed of 4 tables: 2 describing property listings and 2 describing transactions. Properties_2016 and Properties_2017 each consisted of 2,985,217 property listings described by 55 variables, aggregating to a total of 5,970,434 records (full table available in configuration manual). Transactions_16 consisted of 90,275 transaction records and Transactions_17 comprised an additional 77,613 sales transactions (167,838 sales total), corresponding to the Properties tables. Sale prices were not available for each transaction, instead, and for the sake of privacy and security, log errors produced by Zillow's house price prediction model accompanied each transaction. This log-error value is a proxy for house price, as it can be back-translated to discrete price values using the Zillow 'Zestimate' model.

3.2 Pre-Processing and Transformation

Data was uploaded to a personal Google Drive mounted in a Google Colab environment. This allowed for importing of CSV data and use of Google's cloud platform for storage and GPU-assisted analytics. Properties and Transactions were merged via parcel_id such that variables describing the properties were associated with transaction prices and log errors. This yielded an analysis dataframe comprised of 167,838 sales 60 associated variables.

Categorical variables coded using numbers were converted to categorical types so as to avoid incorrect treating coded variables as numeric data. Variables stored as Int64 and Float64 were converted to int16 and float 16 respectively, so as to reduce computational

speed and expense. These variables were not sufficiently granular to require memory-intensive data types. Categorical variables with especially large number of classes were removed due to computational constraints. For example, the census and census_raw variables had an excessively large number of classes within, such that dummy encoding turned the dataframe from a 190-column matrix, to a matrix with 40,000 columns. This yielded prohibitive computational costs. However, the removal of these variables were justified by analysing model performance with and without these variables included no significant impact of removal was found.

Variables were investigated for missing values. "NaN" values for variables such as "garden perimeter" were interpreted as the apartment lacking a garden, rather than the data being missing. In instances where "build date" and "assessment date" were NaN, the median was imputed. Other variables missing large proportions of values were not imputed for and were not removed. This was because the exercise of the GA in this analysis is to determine which features deliver the most value through inclusion; the GA might find that inclusion of a variable missing in 90

3.3 Feature Selection and Engineering

Feature Selection for this analysis was minimal, because the key aim of this investigation was to uncover novel methods of automated feature selection, through the use of GA. However, some features were removed before any optimisation was performed, due to their being identical and redundant with other features. For example, Variables such as Zones, Zipcodes, hood and city were removed due to their containing redundant data.

This section comprised completion of Objective 2 (See Section 1.1, Table 1).

4 Design Specification

The design of this investigation is comprised of a three-tier system (Figure 2), in which the inputs, outputs and process flow of the analysis are presented. The Data Layer includes the processes by which data are sourced and the locations in which they are stored (Google Drive and Excel in this instance). The Processing layer describes the process flow through which the data traverses and is deferentially transformed, analysed, and presented depending on the analysis component. As can be seen from the Processing Layer, the construction of the novel biologically-inspired GA variants are sequential and modular; this will be expounded in the Implementation section. The Exploratory Analysis, Modelling & Machine Learning and Visualisation processes in the processing are the sources of the investigation's outputs. These outputs are represented in the Output Layer, and include novel insights which may inform future analyses, as well as predictive and optimization models ready for application to new datasets.

Two novel biologically-inspired Genetic Algorithm methodologies were investigated in this analysis. These are the Co-Location Genetic Algorithm, and the Multi-Chromosomal Genetic Algorithm.

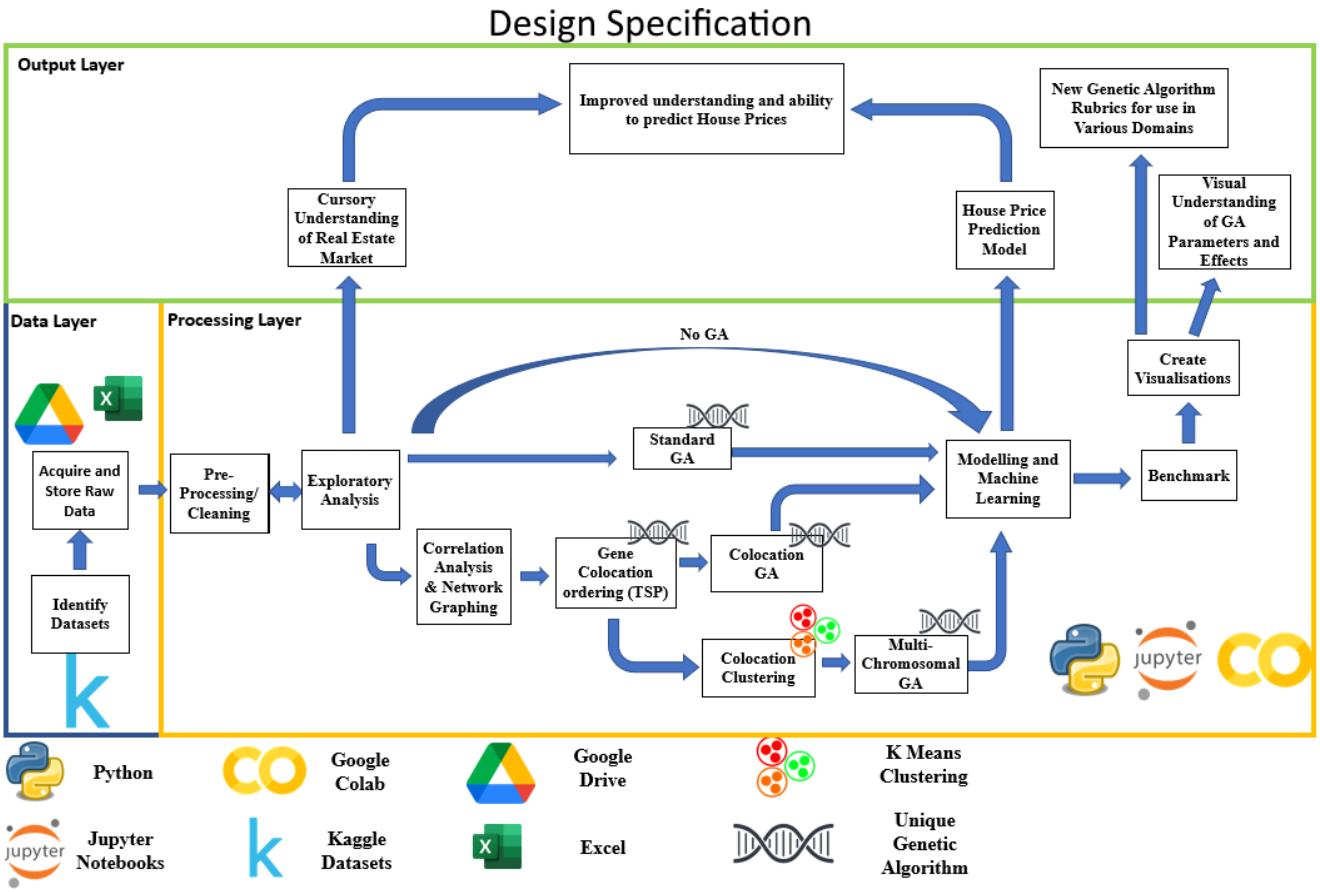


Figure 2: *Design Specification for the Multi-Chromosomal and Co-location Genetic Algorithm investigation.*

Co-Location Genetic Algorithm:
The Co-location Genetic Algorithm (CLGA) will require two preliminary steps. A correlation analysis will assess the relatedness of features, which can then be converted into network graph. Minimising the route through this graph is a Travelling Salesman Problem (TSP), which can be solved using a Genetic Algorithm with a partial-map crossover operator (Hussain et al.; 2017). The optimal route determines the order of genes on the chromosome (Figure 3). This models natural biology, in which related and co-operative genes associate beside one another on chromosomes. Following the optimised ordering of genes on the chromosome, the CLGA follows the same process as the SGA.

Multi-Chromosomal Genetic Algorithm:
The Multi-Chromosomal Genetic Algorithm (MCGA) follows a similar rubric to the CLGA, with additional pre-processing. Following the co-location of genes, the network is translated onto a 2D plane, to facilitate K Means Clustering. Plots comparing the inter-cluster distance and K, will assist in identifying an appropriate K. Each cluster specifies a sub-chromosome, onto which component genes are assigned. These sub-chromosomes crossover with their sub-chromosomal pair, and sub-chromosomes are recombined as one chromosome during the modelling phase (Figure 4). This allows each sub-chromosome to evolve at different rates, which will allow the GA to speed up resolution of more difficult components of the problem, without jeopardising *solved* components.

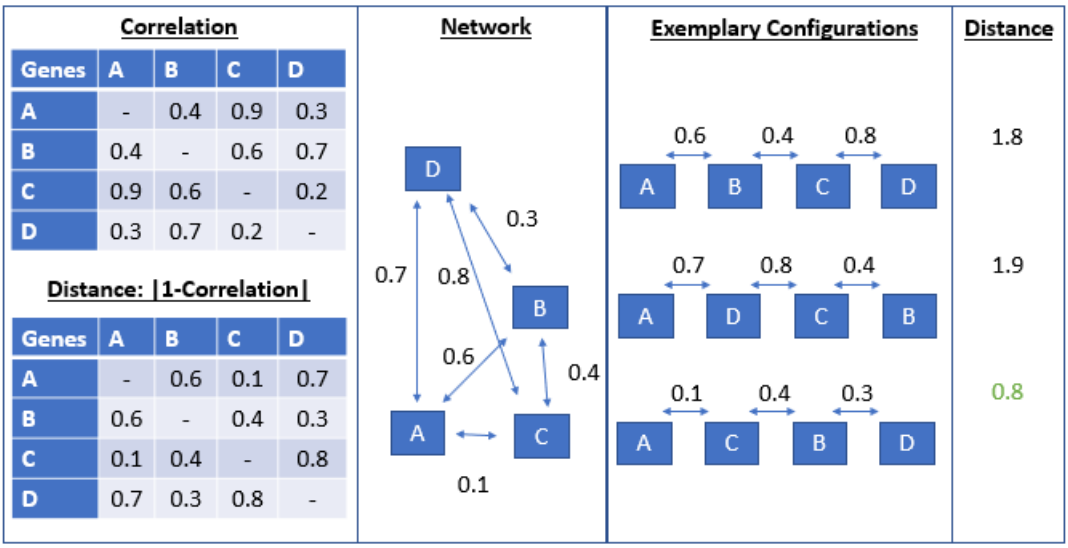


Figure 3: *Conceptualisation and rationale of Co-Location Genetic Algorithm*

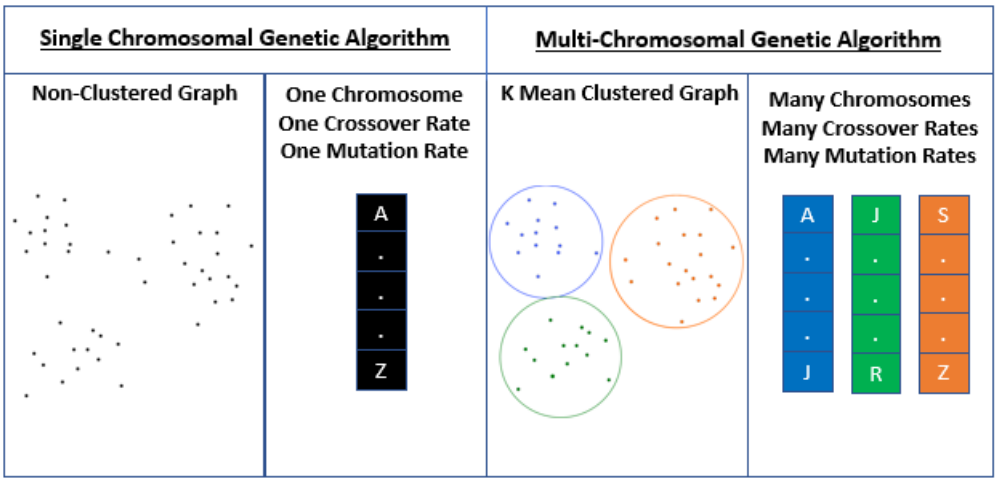


Figure 4: *Conceptualisation and rationale of Multi-Chromosomal Genetic Algorithm.*

5 Implementation, Results and Evaluation

In this section, each of 6 primary experiments comprising this study are separated into subsections. Within each experiment are sub-experiments. The implementation of each sub-experiment is given first, followed by a presentation of the numeric and graphical results. Based on the results, actionable interpretations and recommendations are given in the evaluation components.

5.1 Exploratory Analysis

An exploratory analysis was performed to better understand the data being worked with in this investigation. This analysis was aimed at providing insights into the dataset, and prediction of property prices, that would be useful for improving future data collection and price prediction activities.

5.1.1 Descriptive Analysis of Target Variable

Implementation:
To better understand how the target variable (predictor error) is distributed, descriptive statistics (Table 3) and histograms of the target variable were created (Figure 5). Due to the especially wide range, and high kurtosis, additional histograms were generated to zoom into those values that fell within the interquartile range.

Results:
As can be seen from the histograms (Figure 5) and descriptive statistics (Table 3), the distribution of errors are especially tight around the median (approximately 0). The high kurtosis is exemplified by the high range (9.918), and especially low interquartile range (0.065) and standard deviations (0.166). There is a slight positive skew, as indicated by the mean being slightly larger than the median.

Table 3 - Summary Statistics

Count	Mean	Standard Deviation	Min	Max	25^{th} Percentile	Median	75^{th} Percentile	Interquartile Range	Range
167,888	0.014	0.166	-4.656	5.263	-0.025	0.006	0.039	0.065	9.918

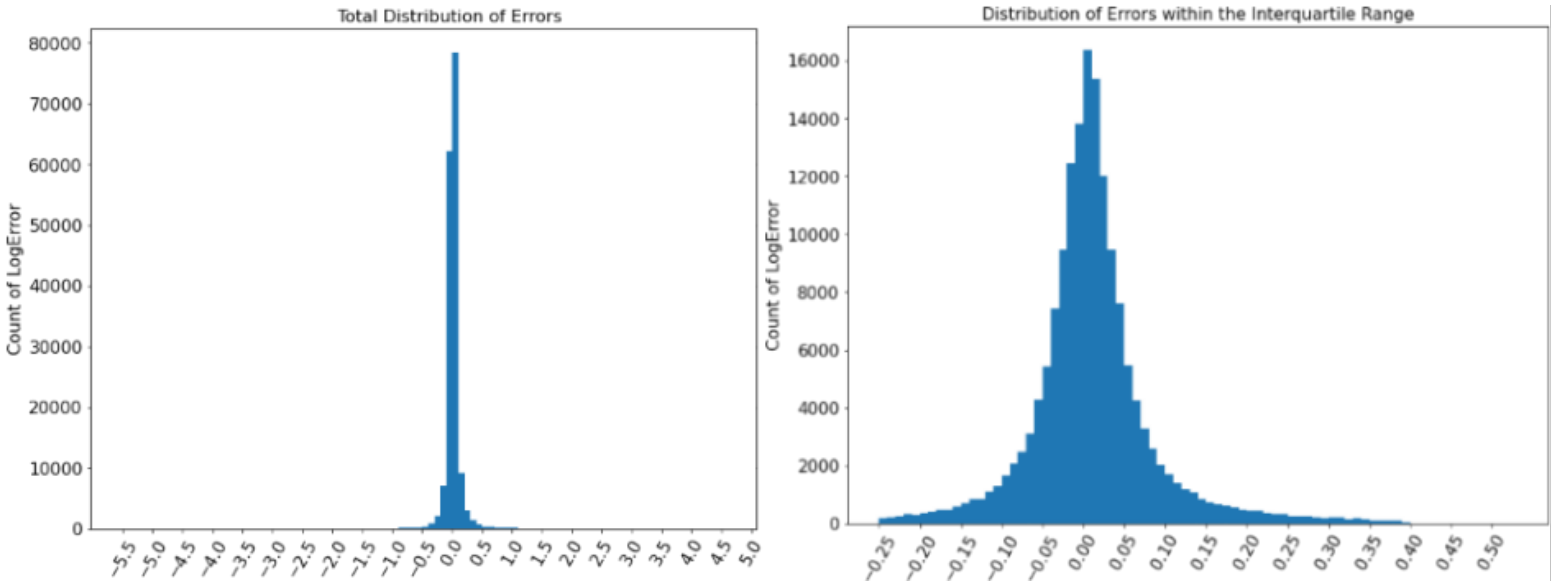


Figure 5: *Total distribution (left) of all prediction errors across both 2016 and 2017. This shows an extremely high kurtosis, with almost all values falling near 0. Distribution of prediction errors that fall within the interquartile range (right). These appear closer to a normal distribution, but still exhibit high kurtosis.*

Evaluation:
Given that the majority of values tend very closely toward 0, this experiment demonstrates that the Zestimate predictive model is highly accurate. When the observations within the interquartile range are examined, it can be seen that the values approach normality around a median of 0.006. However, the slight disparity between median and mean indicates that this distribution has a slight bias toward the positive tail, indicating that the model is more likely to overestimate price predictions than underestimate; this

is a key insight that could improve the prediction model. Finally, although the model is typically accurate, the extreme disparity between range and interquartile range highlights that the Zestimate model can be extremely wrong in some instances. Together, these insights would motivate further inspection of those outlier errors on the fringe. It may be that these errors are caused by missing values in key features.

5.1.2 Univariate Correlations and Associations with Target Variable

Implementation:
To better understand how each feature in the dataset associates or is correlated with the target variable, correlation and/or analysis of variance tests were performed. Before these statistical tests were performed however, the normality of the target variable needed to be tested. A Kolmogorov-Smirnov test compared the distribution of the logerror variable to a standard normal distribution. This found the distribution of the target variable to be significantly different to the normal distribution ($KS = 0.4$, $p < 0.001$). This required that subsequent statistical tests use non-parametric alternatives to the parametric tests that assume normality.

For correlation of numeric features and binary categorical variables to the target variable, Spearman's Rho was used.

To understand the association of categorical variables with multiple factors with the target variable, the Kruskal Wallis test was performed. Where significant differences were found between the factors, post-hoc Dunn tests were performed. Bonferroni correction was used to account for the effect of multiple comparisons on statistical significance.

Results:
Due to there being 52 predictive variables, it is not appropriate to include all visualisations and statistical tests here. Instead, those correlation analyses that yielded the strongest associations are included. Also, those Kruskal Wallis tests that achieved significance and demonstrated associations with the target variable were included (Figure 6).

A significant correlation was found between the total living area of the property and the error of the prediction, however, the low positive Rho (0.067, $p < 0.001$) indicates that there is only a small positive association.

A Kruskal Wallis test found that there was a significant interaction between the type of aircon system and the propensity for incorrect prediction of house price ($KW = 39.1$, $p < 0.001$). Post-hoc Dunn tests found that the differences existed between Aircon system 1 and Aircon System 13 ($p < 0.001$) and Aircon System 1 and nan ($p < 0.001$) (data was missing) and Aircon system 13 and nan ($p < 0.001$). These associations survived Bonferroni correction. Significant interactions between the FIPS (area) ($KW = 76.3$, $p < 0.001$) and county ($KW = 76.3$, $p < 0.001$) codes of the property and the propensity for incorrect prediction of house price were also found. Post-hoc Dunn tests found that the differences existed between FIPS code 6037 and 6059 ($p < 0.001$), 6037 and 6111 ($p < 0.001$), and county codes 3101 and 1286 ($p < 0.001$), and 3101 and 2016 ($p < 0.001$). These associations survived Bonferroni correction.

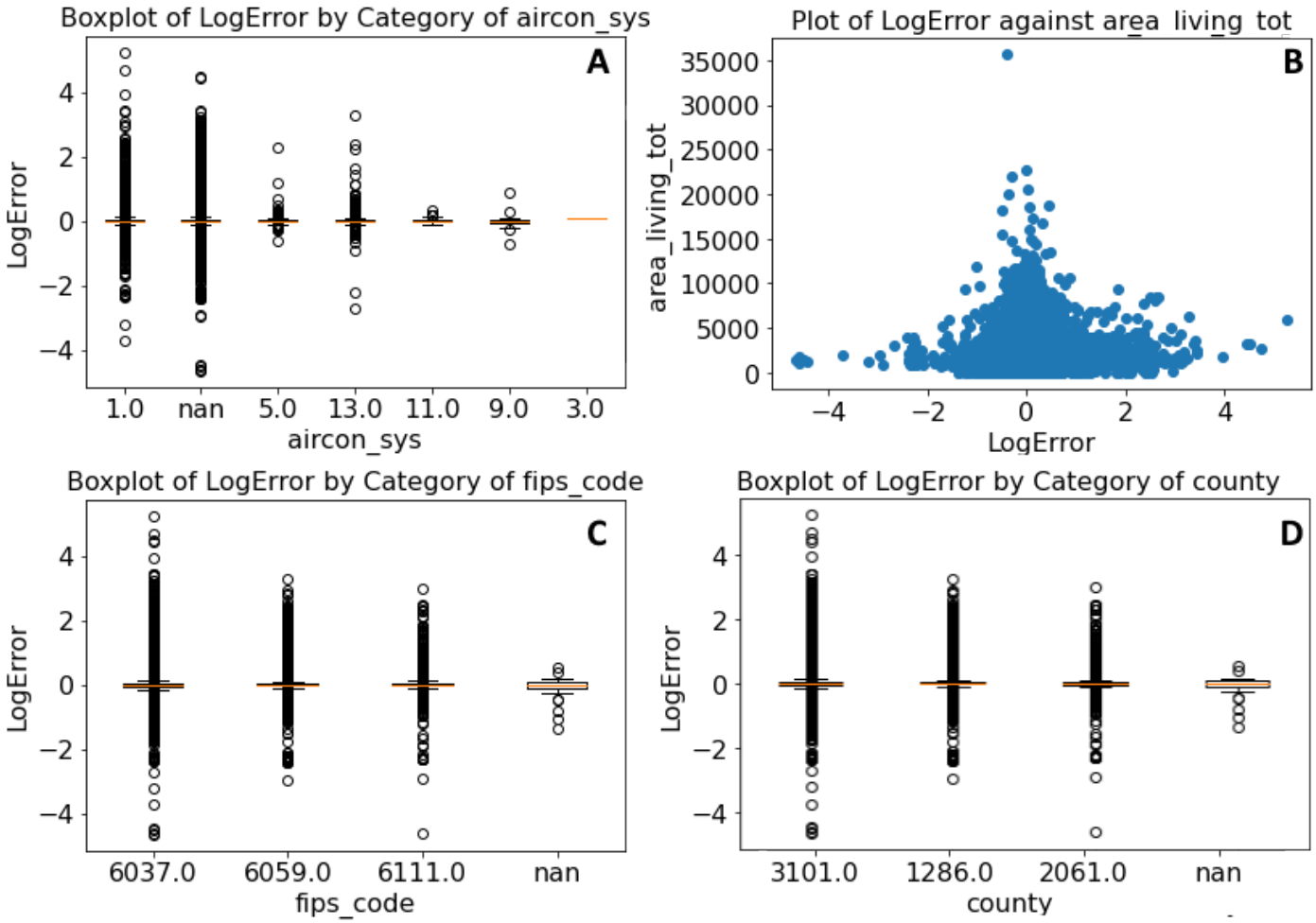


Figure 6: *Kruskal Wallis tests demonstrating significant associations between aircon system* ***(A)****, FIPS code* ***(C)*** *and County* ***(D)*** *with price prediction error. Spearman's Rho correlation found there to be a weak positive correlation between prediction error and the total living area in the property* ***(B)****.*

Evaluation:
The significant associations between the living area with the log error target variable was weak but positive. This would suggest that properties with larger living areas are more difficult to price. Although the associations in the aircon system analysis found there to be significant differences between each category, the differences were minimal. However, a significant finding of this test was the significant difference between the missing and non-missing factors. This indicates that missing data in this column has a significant effect on prediction accuracy, as compared to other factors. If the values were missing at random, they would not throw significant differences. Although the associations between the FIPS and count codes with the prediction error are also slight but significant, this test revealed that the variance of prediction error differs substantially between FIPS and county codes. This indicates that it is more difficult to predict property prices in some FIPS code and county areas, compared to others.

5.1.3 Analysis of Prediction Error as a Function of Time

Implementation:
To investigate whether there is a relationship between the time of transaction and the error of price prediction, the monthly median error of transaction price was plotted against time. To further investigate whether error was impacted by the sample of available properties, the number of transactions occurring per month was plotted against time.

Results:

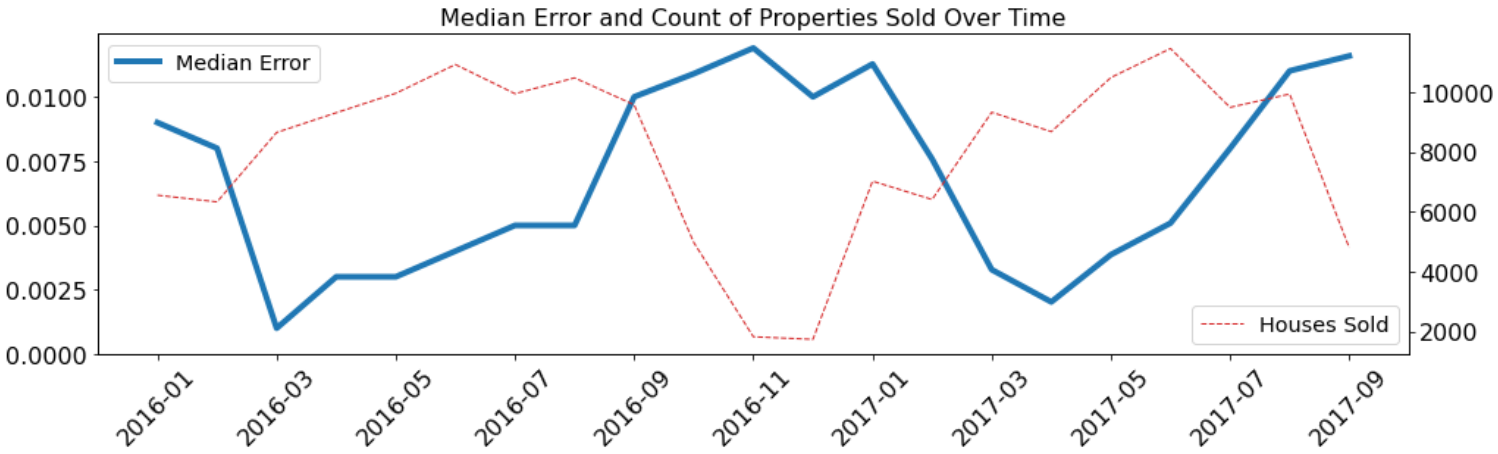


Figure 7: *A graph of median house price prediction error and count of property transactions.*

Evaluation:
This visualisation (Figure 7) indicated that house prices appear to follow a seasonal trend, with the spring months of each year associating with higher price prediction errors. The median error appears to negatively correlate with the number of properties on the market. This might be due to the lesser availability of transactions on which to train the model, leading to greater error. However, there are points at which the trend lines intercept, such as in 2016-09, when the number of transactions was high, yet the error was also high. One possible explanation for this is the dynamics of supply and demand, which is not captured directly in the dataset. That is, during the "flatter" periods of the graph, the predictive model focuses on the qualities and characteristics of the property. However, as the number of properties on the market decreases, demand for properties increases relative to supply, and drives up the price. This effect of supply and demand are not captured in the characteristics of the house and may therefore not be incorporated into the Zestimate's predictions, leading to increased error. This would motivate incorporation of an additional feature that summarized the trend in number of properties on the market.

5.1.4 Analysis of Geographic Location's on Prediction Error

Implementation:
To understand how price prediction error differs with geography, each transaction was mapped using the latitude and longitude values. This was superimposed on a map of Los Angeles. Error values determined the dot colour; larger areas were darker shades of blue.

Results:

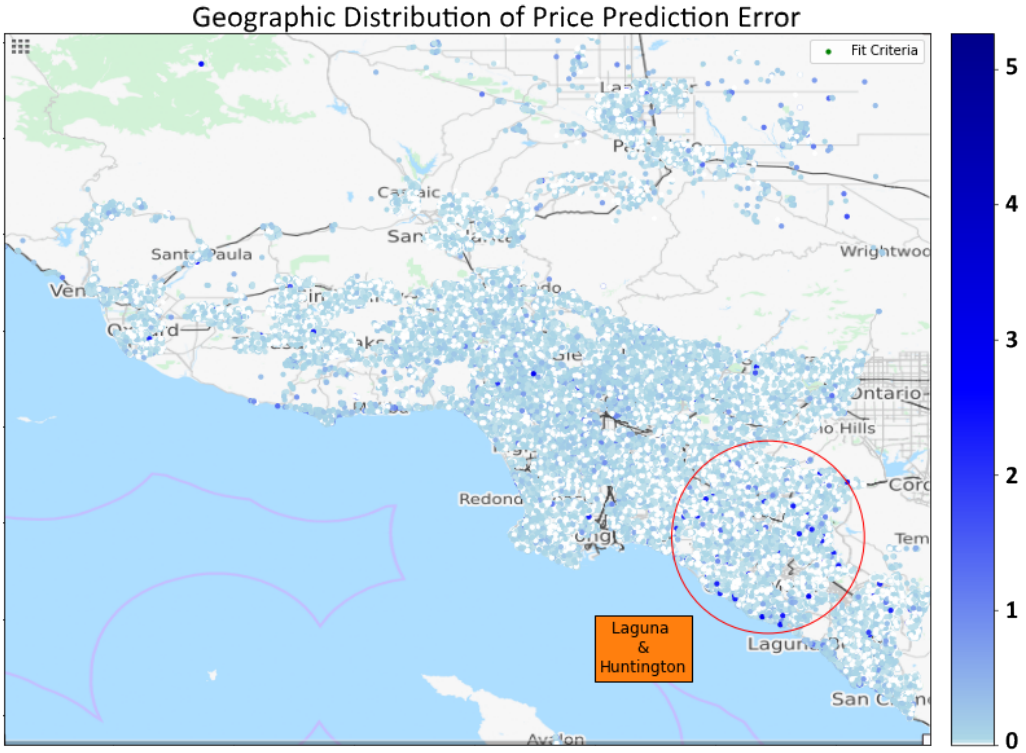


Figure 8: *Display of the geographic distribution of property price predictions. Although price prediction errors can be found throughout Los Angeles, a higher density of errors occur near Laguna and Huntington (Circled Red)*

Evaluation:
This visualisation (Figure 8) demonstrated that prediction errors occur across the total geography. However, there is an area in which prediction errors are apparently denser. The area between Laguna Beach and Huntington Beach (circled in red), appears to have a greater density of transactions with substantial price prediction errors, as indicated by the deep blue dots. This suggests that these areas contain properties that are less amenable to prediction. Possibly due to the unique characteristics of properties in that area. Alternatively, data collection and/or real estate practices might mean that particular features are not collected as faithfully in this location.

Together, this exploratory analysis consisting of 4 sub-experiments, comprised completion of Objective 3 (See Section 1.1, Table 1).

5.2 Unoptimized Modelling

Implementation:
Data mining methods that have found success in previous analyses of price prediction were assessed
The XGBoost (XGB) model, Linear Regression (LR), Decision Tree Regression (DT), Random Forest Regressor (RF) and four variants of Support Vector Regression (SVR) (Linear, Polynomial, Radial Basis Function, Sigmoid) were analysed.

K-Fold Cross-Validation was used to reduce the incidence of overfitting to training data. This process involves the splitting of the dataset to produce different train and test splits for each iteration. A K of 5 was chosen for this study, based on the relatively large number of models being assessed, and the expected computational cost of analysing a large dataset. A K of 5 meant that each model was analysed 5 times, using a different 80% of the data to train the model and a different 20% for testing and validation each iteration.

These unoptimized models were primarily compared to one another using the Mean Absolute Error metric. This metric computes the average difference between the prediction and the actual value, calculated as a percentage of the actual value; an MAE of 0.1 indicates a 10% difference between the prediction and the actual value. they were also evaluated based on the time requirement, because models that required too much time would not be amenable to GA optimisation, due to the inherent iterative nature of the optimisation method.

Results:
After an initial trial, it was apparent that the SVR variants took a prohibitive length of time (over 1 hour each). They were therefore excluded from the results and subsequent analyses, as it was not feasible to wait for their completion.

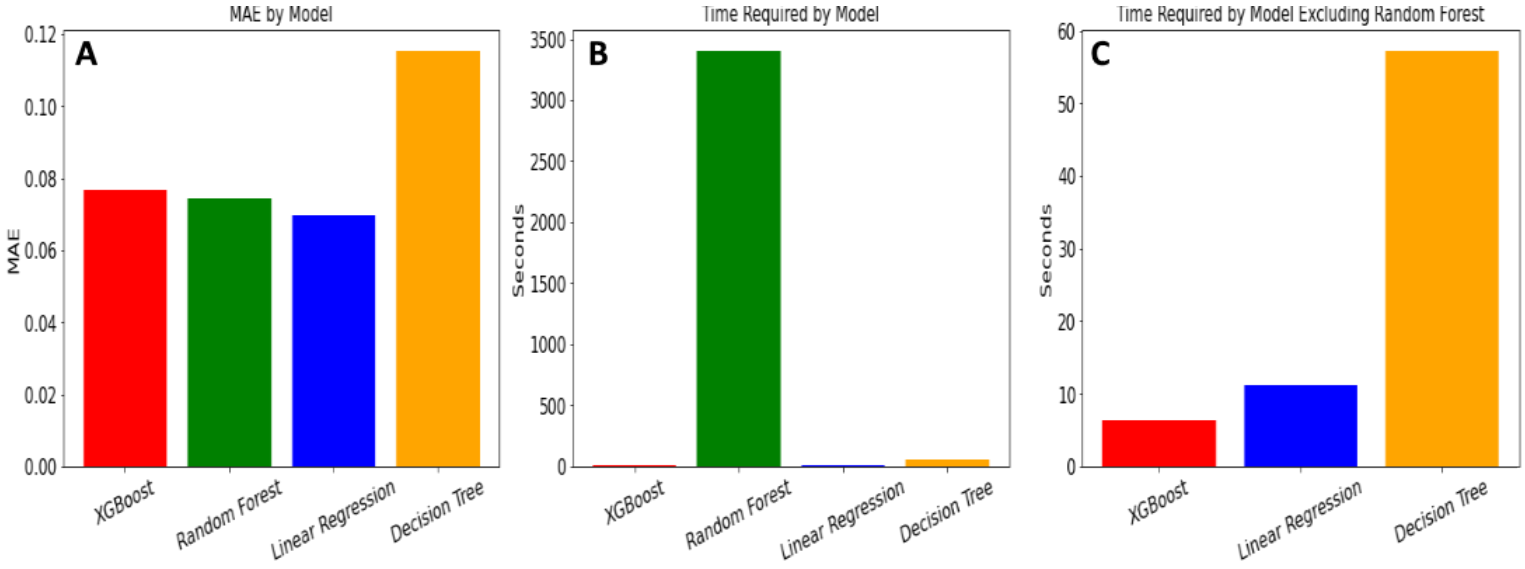


Figure 9: *Graph of results of prediction of house price prediction error by machine learning models not optimized by any genetic algorithm method. A: Best MAE achieved, B: Time required to complete modelling, C: Time required to complete modelling with Random Forest Excluded.*

From the analysis of the XGB, RF, LR and DT models, it was found that the DT was the worst performer, achieving an MAE of 0.115, whereas LR was the best performer, achieving the lowest MAE of 0.0696 (Figure 9 A). RF was found to outperform the XGB Model in terms of MAE (0.0745 vs 0.0767), but as can be seen in Figure 9 B, the RF took far longer than any other model (55 minutes). Figure 9 C shows the comparison of times required for each model without RF included.

Evaluation:
The necessity to exclude SVR variants may have been due to the number of columns involved in this analysis. Following creation of dummy columns, the total feature set can extend to over 4000 columns. Given that SVR are sensitive to dimensionality, they were unable to complete the modelling task in a reasonable length of time.

These results show that the model best able to predict errors in the Zestimate model is LR. The time requirements also inform on how feasible GA optimization each model is. Figure 9 B shows that RF would be impossible to optimize using GA, as each generation would require 55 minutes to run. Conversely, the time requirements indicate that the XGB and LR models are most amenable to optimization by GA. Given that XGB requires roughly half the time as LR, XGB could theoretically achieve twice as many generations of evolution in the same time as LR. Therefore, both are worthwhile for further investigation. Because DTs had been found to be successful in past analyses, and because the time requirement was not prohibitive, the model was also optimized via GA in later experiments.

Together, this subsection comprised completion of Objective 4 (See Section 1.1, Table 1).

5.3 Standard Genetic Algorithm Optimisation of Modelling

Implementation:
Chromosome Structure
The chromosome was composed of an array of 0's and 1's of length equal to the number of features in the dataset (52). Each position in this array mapped back to the features such that a 0 or a 1 corresponded to exclusion or inclusion of a feature, respectively, for that iteration of modelling.
Fitness Function
Fitness for each chromosome was evaluated using three metrics. Chromosomes were primarily evaluated based on their MAE (MAE_1) score. However, to reward chromosomes that required fewer features, a second MAE metric (MAE_2) was used. MAE_2 equalled MAE_1 rounded to 2 decimal places. When two chromosomes were equal in MAE_2, the chromosome that required fewer features was ranked higher. Conversely, if two chromosomes achieved equal MAE_2, and used an equal number of features, the chromosome with higher MAE_1 was superior.
Initialisation
To initialise the population, 20 arrays were generated such that each element had a 50:50 chance of being 1 or 0.
Selection
Following each generation, selection probabilities (P_s) were assigned to each chromosome. This was calculated based on the MAE achieved by the chromosome when that set of features was applied to the modelling task. This made it more likely that the best performing chromosomes were selected for reproduction.
Mutation Recombination
Two chromosomes were selected based on their selection probability, described above. Each bit of each chromosome had an equal probability of mutating (flipping from 0 to 1 or 1 to 0). Following this round of mutation, chromosomes were crossed over at random points. This was to allow the best genes of parent chromosomes to combine and

persist in the population, and the worst genes to be removed. Following recombination, both parents and both children were returned to the population. The worst performer of the population was removed each generation. This elitism method ensured the best chromosomes survived through each generation and that the worst genes were removed.

Results:

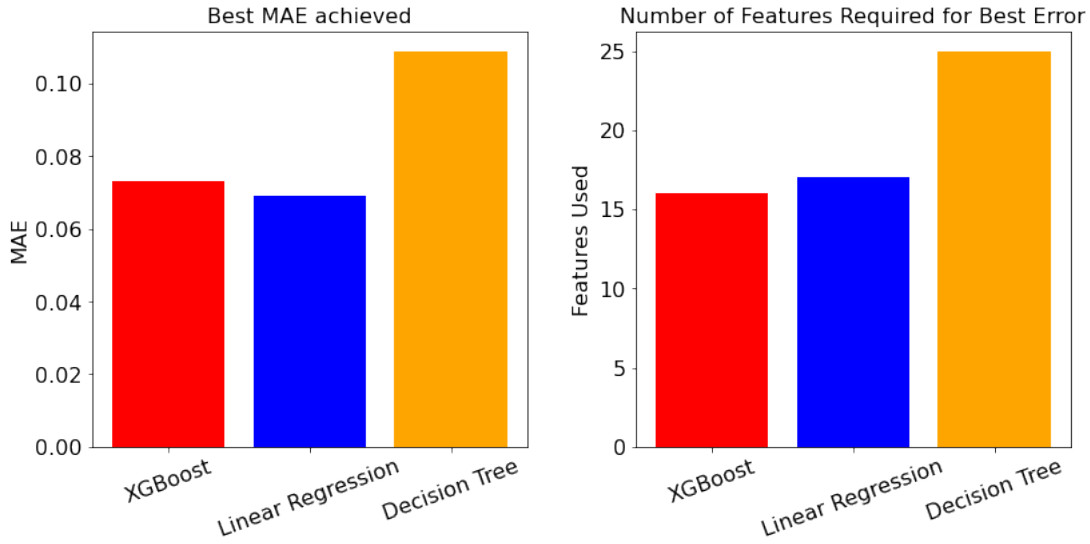


Figure 10: *Chart of best MAE achieved (left) and fewest features required (right) by each model following 250 generations of feature selection optimisation by Standard Genetic Algorithm*

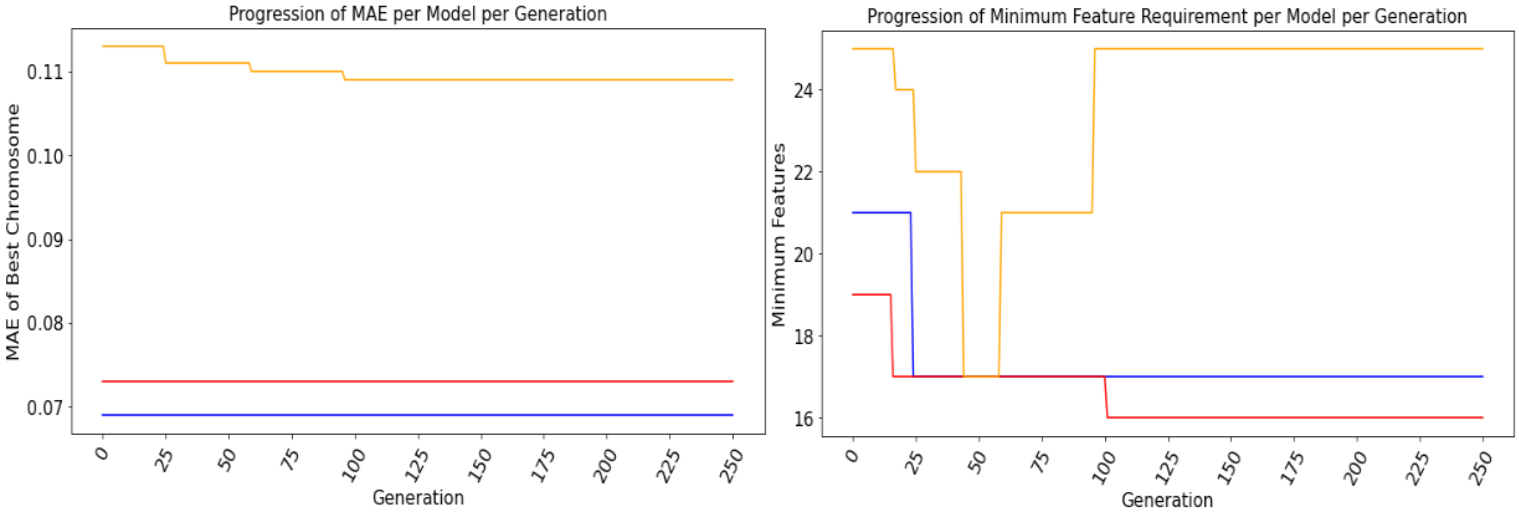


Figure 11: *Chart of the progressive improvement in MAE and reduction in number of features required by each model through the course of 250 generations of Standard Genetic Algorithm optimisation.*

There was no substantial improvement in the Best MAE achieved by each model following optimisation of feature selection by an SGA, except for in the DT, which exhibited

a 10% improvement from 0.115 to 0.109 MAE (Figure 10, left)). XGB and LR improved from 0.0761 and 0.0696 to 0.073 and 0.069 respectively.

Fewer features were required to achieve better results in the XGB (16) and LR (18) models compared to the DT (25) (Figure 10, right).

Model error was not improved for the XGB and LR models over the course of evolution, but the DT exhibited some improvement (Figure 11, left), that plateaued after 100 generations. For XGB and LR, the number of features required to maintain their performance reduced with each generation, from 22 and 19, to 17 and 15 respectively (Figure 11, right).

Evaluation:
Although there was no substantial movement in the MAE for the better performing XGB and LR models, the SGA optimised the modelling process such that fewer features were required to achieve the same performance. This means that future analyses will not need to include as many erroneous features that contribute no predictive power but may cause overfitting of data.

This subsection comprised completion of Objective 5 (See Section 1.1, Table 1).

5.4 Construction and Deployment of Co-Location Genetic Algorithm

In order to construct the CLGA, in which genes are ordered on the chromosomes such that the most related genes are close to one another, two steps were required. A network graph was first constructed based on correlation values. The optimal route through this network was then used to determine the optimal order of genes on the chromosome. Following gene order optimisation, the new gene order was applied using the SGA method.

5.4.1 Network Construction and Travelling Salesman Genetic Algorithm

Implementation:
Binary class categorical variables were represented as 0s and 1s and were combined with the numeric variables. Multi-class categorical variables were included as dummy variables. The dataframe was then used to compute a correlation matrix. This was converted to a distance graph by subtracting each of the correlation values from 1. Each distance was used to calculate the relative position of each variable in a fully connected network.

The shortest route through the network would yield the order of genes which minimised distance between related genes on the chromosome. This is a Travelling Salesman Problem (TSP), which finds the shortest route through a network. An SGA, was deployed to this effect, but each element represents a point/stop in the route rather than representing inclusion/exclusion of a feature. Fitness depended on the shortest route length (sum of distances from each gene in the chromosome to the next). Finally, the partial map crossover operation was used to ensure no two nodes on the graph could appear twice following recombination. This method is most effective in Travelling Salesman Genetic Algorithms (Hussain et al.; 2017).

Results:
Given that each predictor in the network graph (Figure 12) has a correlation and there-

fore a distance relationship with every other predictor, this network is fully connected. Larger distances indicated smaller correlations.

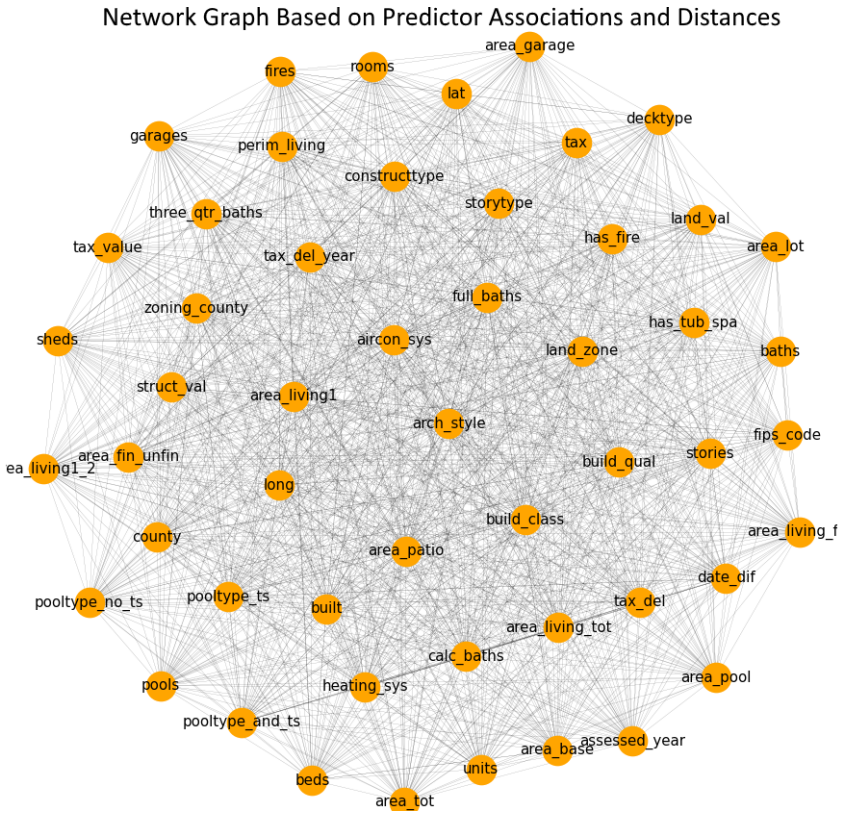


Figure 12: *Network representing correlations (grey lines) between each feature (nodes).*

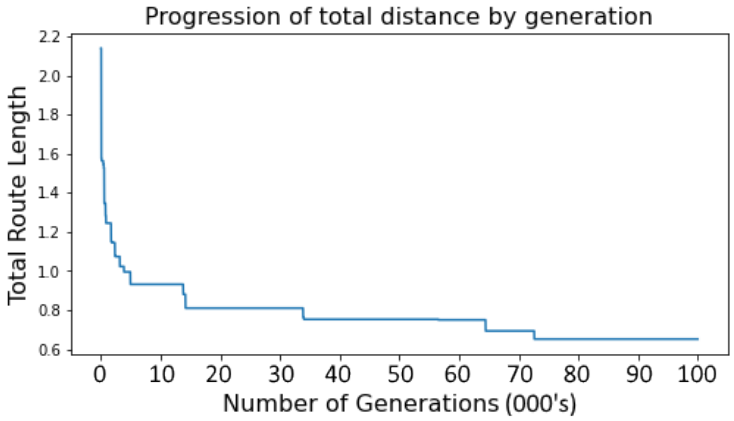


Figure 13: *Minimisation of the total network route distance.*

Following 100,000 generations, the Travelling Salesman Genetic Algorithm (TSGA) reduced the total distance required to traverse the network by 68% (Figure 13).

Evaluation:
Although the network graph was generated as a precursor to the TSGA, some insights may be garnered from the visualisation. Because nodes that are close on the graph are those most correlated with one another, it can be understood from the graph that the

total area of the property (area_tot), the number of bedrooms (beds), the number of units (units) and the base area (area_base) of the property are strongly correlated (bottom 4 nodes). This is unsurprising, as larger properties would have larger areas and rooms. However, among others, there are interesting correlations between the air conditioning system (aircon_sys), architecture (arch_style) and building quality (build_qual). This correlation in the centre of the network may suggest that particular architectural styles are each supported by typical/standard air conditioning systems, and that some styles are of lower quality than others.

For the TSGA optimization, it appears as though the majority of the improvement achieved through evolution occurs within the first 10,000 generations. Over this span, the total distance improves from 2.1 to 0.8 (62%). Considering that the following 90,000 generations only yield an additional 6% improvement compared to the baseline, this exercise demonstrates clear diminishing returns. For more demanding TSGA experiments, fewer generations may be more efficient.

5.4.2 Co-Location Genetic Algorithm Optimisation of Modelling

Implementation:
Once the optimal order of genes on the chromosome was derived from the TSGA, the analysis dataset was re-indexed such that it matched this optimal order. This co-location dataset was then fed through the SGA methodology as before.

Results:

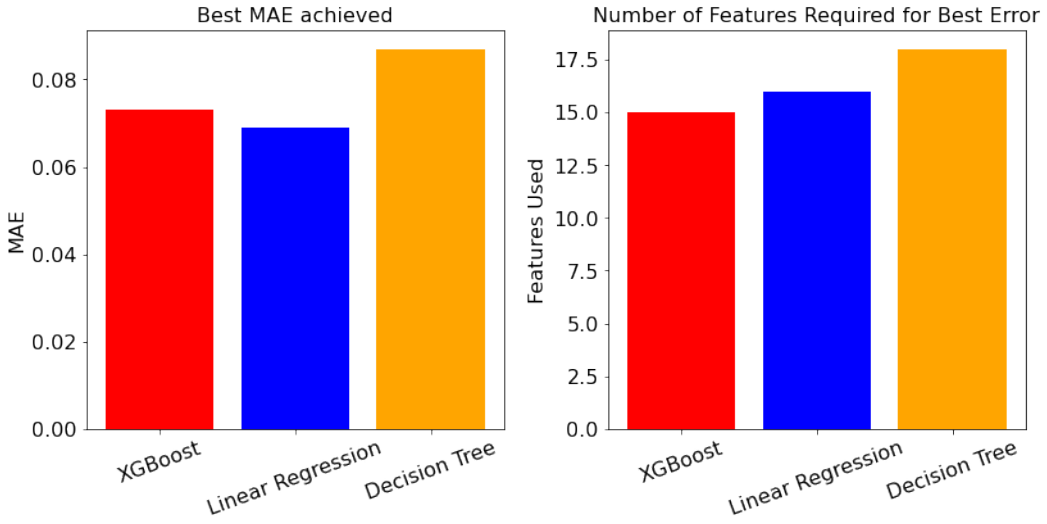


Figure 14: *Chart of best MAE achieved (left) and fewest features required (right) by each model following 250 generations of feature selection optimisation by Co-Location Genetic Algorithm*

There was no significant improvement in the Best MAE achieved by the LR and XGB models compared to the SGA; best MAEs remained as 0.073 for XGB and 0.069 for LR. However, the DT showed an additional improvement from 0.109 to 0.087 (Figure 14, left).

Again, fewer features were required for each model, with XGB achieving a new minimum of 15 features, LR achieving 16, and DT achieving 18 (Figure 14, right).

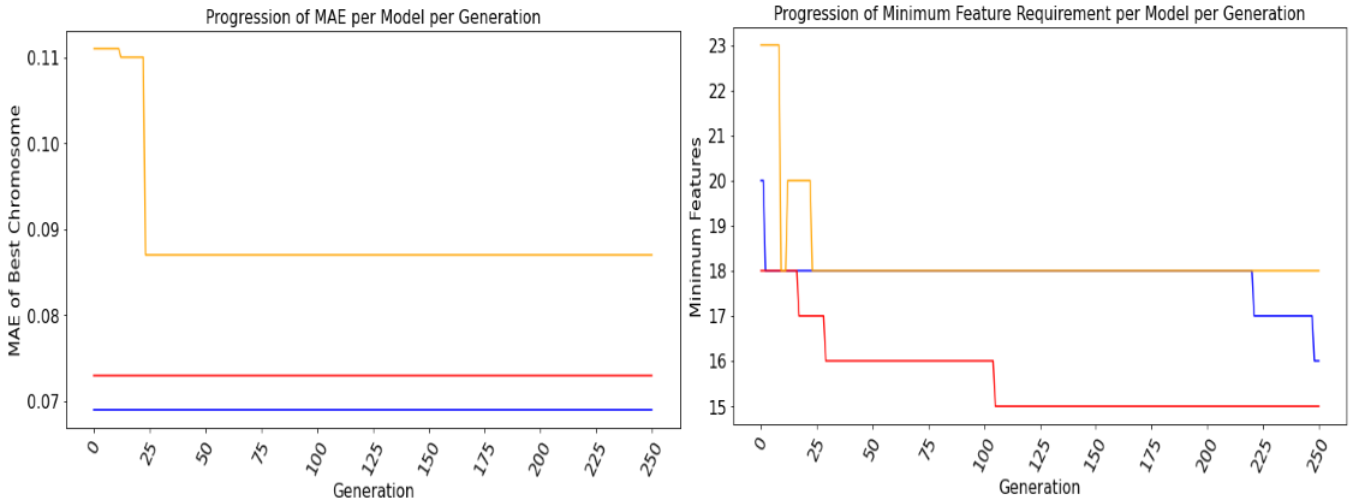


Figure 15: *Chart of the progressive improvement in MAE and reduction in number of features required by each model through the course of 250 generations of Co-Location Genetic Algorithm optimisation.*

Model error was not improved for the XGB and LR models over the course of evolution, but the DT exhibited a sharp improvement after the 25th generation (Figure 15, left). The evolutionary reduction of features required for XGB and LR was more gradual and maintain (Figure 14, right).

Evaluation:
From these results it would be suggested that the Co-location processing step improves the rate at which each model evolves in terms of features required. The steeper drops toward fewer features indicates that the algorithm is more quickly able to eliminate valueless features from the set.

Again, the generations were not accompanied with improvements in the XGB and LR models, but the DT appears to be more amenable to significant improvement. The plateau in the DT, in terms of both MAE and Features required may indicate that the algorithm fell into a local minimum and was unable to escape. A larger mutation rate may help avoid this scenario in future investigations.

5.5 Construction and Deployment of Multi-Chromosomal Genetic Algorithm

As with the CLGA, construction of the Multi-Chromosomal algorithm required a pre-step. The network created during the co-location process was then translated onto a 2D plane such that K Means clustering was possible. Genes separated into clusters were later separated into sub-chromosomes during the multi-chromosomal algorithm optimisation stage.

In this section, the correlation analysis and construction of the Network Graph completed sub-objective 6a, the implementation of a Travelling Salesman Genetic Algorithm completed sub-objective 6b, and the implementation and evaluation of the Co-Location Genetic Algorithm-mediated optimisation of machine learning completed sub-objective 6c. Together, these steps comprise completion of Objective 6 (See Section 1.1, Table 1).

5.5.1 K Means Clustering for Generation of Multi-Chromosomes

Implementation:
The Co-Location and MCGA comprised an additional step on top of the CLGA process. Namely, the nodes in the graph were clustered using K Means Clustering. This required translation of the distances onto a 2D plan.

The suitable value for K was derived from visual inspection of the network, and with the support of an elbow chart describing the total sum squared distances between clusters with variable K.

Results:

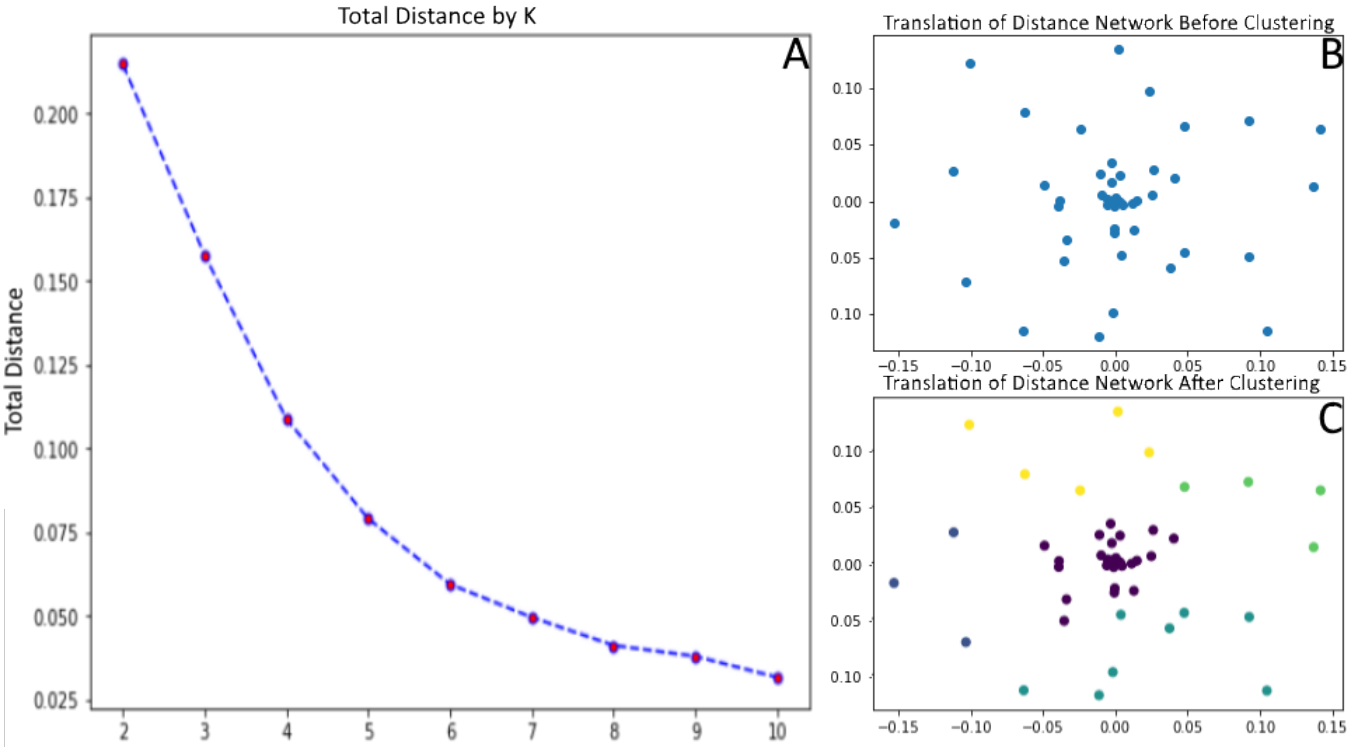


Figure 16: *Elbow plot of Total Sum Squared distance achieved by varying K* ***(A)****. Translation of Distance Network onto 2D plane, allowing for K Means clustering* ***(B)****, each predictor is represented by a point. Predictors are clustered into groups (colour coded), that will partition onto separate sub-chromosomes during the Multi-Chromosomal Genetic Algorithm rubric* ***(C)****.*

Evaluation:
An elbow plot showing reductions in Total Sum Squared Error/Distance between nodes on the plane was used to determine the optimal value for K (Figure 16, A). This plot indicated that 5 clusters may be most beneficial, as the reductions in error were reducing significantly after K=5. As can be seen from Figures 14 B and C, each node is sequestered into clusters that will allow for separation across sub-chromosomes.

5.6 Multi-Chromosomal Genetic Algorithm Optimisation of Modelling

Implementation: Once the genes were separated across chromosomes, the method followed the same rubric as with the SGA. The difference was that instead of each parent crossing over at a single point, the parent was split into its component chromosomes and each sub-chromosome was crossed with the corresponding sub-chromosome of the other parent.

Results:

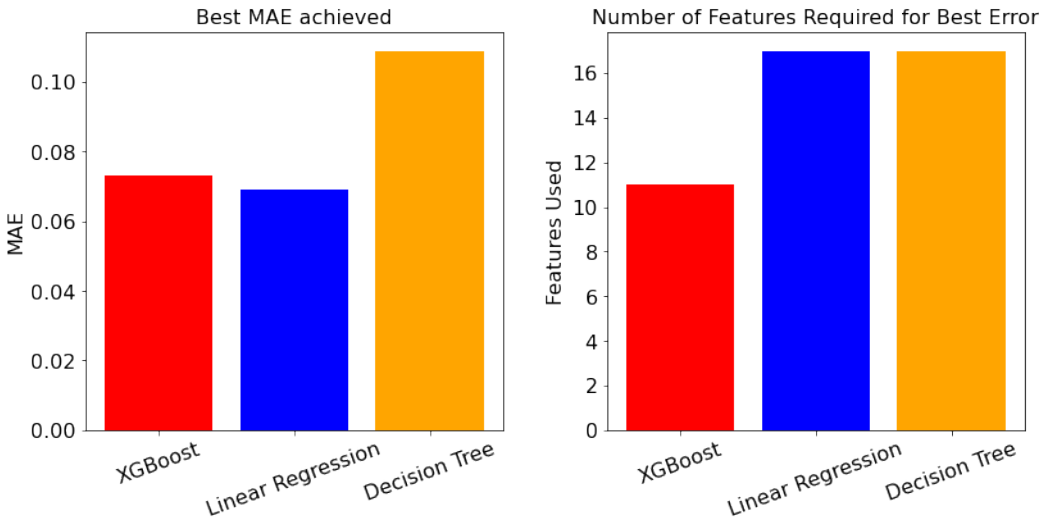


Figure 17: *Chart of best MAE achieved (left) and fewest features required (right) by each model following 250 generations of feature selection optimisation by Multi-Chromosomal Genetic Algorithm*

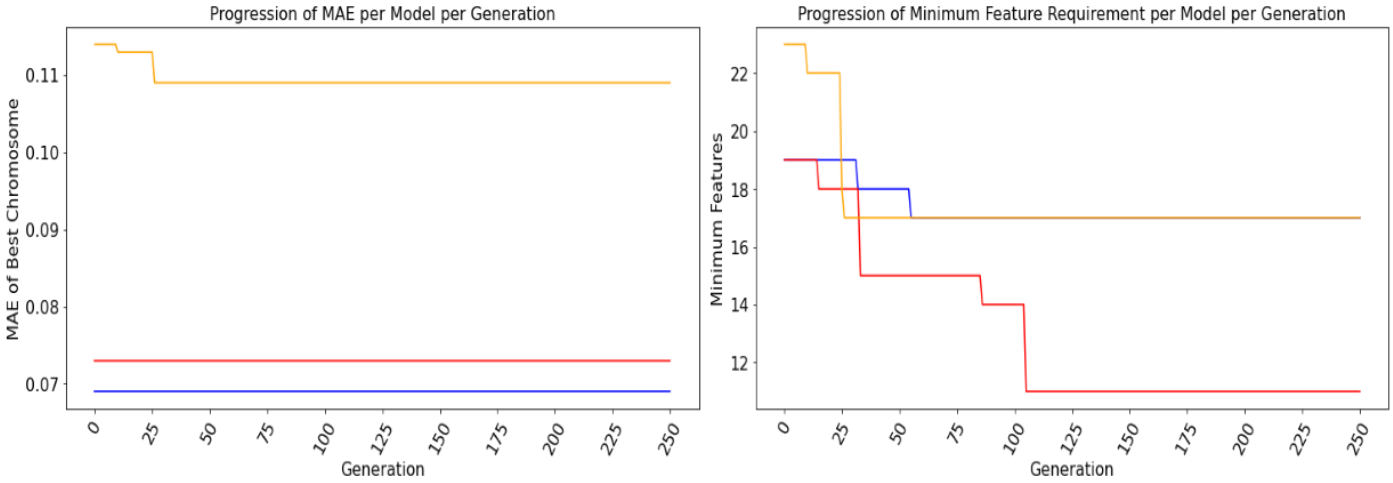


Figure 18: *Chart of the progressive improvement in MAE and reduction in number of features required by each model through the course of 250 generations of Multi-Chromosomal Genetic Algorithm optimisation.*

In terms of MAE, MCGA did not confer any advantage as compared to CLGA,

achieving the exact same values for each model (XGB: 0.073, LR: 0.069) except for DT, in which performance reverted to that achieved by SGA (0.109) (Figure 17, left). The DT MAE did not improve past the 25th generation (Figure 18, left).

Features required by XGB achieved the low of 11, after 100 generations of MCGA. The minimum features required by LR and DT bottomed and plateaued at 17 features, after 25 and 50 generations respectively (Figure 18, right).

Evaluation:
These results suggest that the MCGA optimisation method is least capable of improving MAE but may be most effective in optimising features. There were no improvements to the XGB and LR model MAE, indicating that MCGA is not more capable of improving the predictive power of these models. However, the improvement in the number of features required by XGB is stark. A new minimum number of features required to maintain peak performance was achieved through MCGA. For the other models, in both MAE and Feature evolution, MCGA appeared more prone to early plateau and lack of progress. This may indicate that MCGA is more likely to fall into local minima from which the algorithms are unable to escape. MCGA may be less likely to fall into these local minima if accompanied by more aggressive methods of ensuring diversity and genetic variation, such as through additional injection of new genetic material, or dynamic increases in the mutation rate.

In this section, the Translation of the Network Graph onto a 2D plane and subsequent K Means clustering of genes onto sub-chromosomes comprised completion of sub-objective 7a. The implementation and evaluation of the Multi-Chromosomal Genetic Algorithm-mediated optimisation of machine learning completed sub-objective 7b. Together, these steps comprise completion of Objective 7 (See Section 1.1, Table 1).

5.7 Benchmarking of Genetic Algorithm Optimization Methods

Implementation:
To benchmark all GA based on their ability to optimize feature selection, the best MAE achieved, and the minimum features required by each model were grouped and compared by optimization method (unoptimized, and the three types of GA). This allowed for comparison of the absolute best performances achieved by each model and GA.

The speed and efficiency with which each GA optimized both prediction of price prediction error, and with which feature requirements were minimised. In this comparison, GA that helped their models achieve maximal performances with fewer generations were deemed superior.

Results: Although all GA optimized models outperformed unoptimized versions in terms of MAE, no GA method showed superiority in improving the predictions by XGB or LR. However, the DT was best optimised by the CLGA (0.087) as compared to the SGA (0.109) and MCGA (0.109) (Figure 19, left).

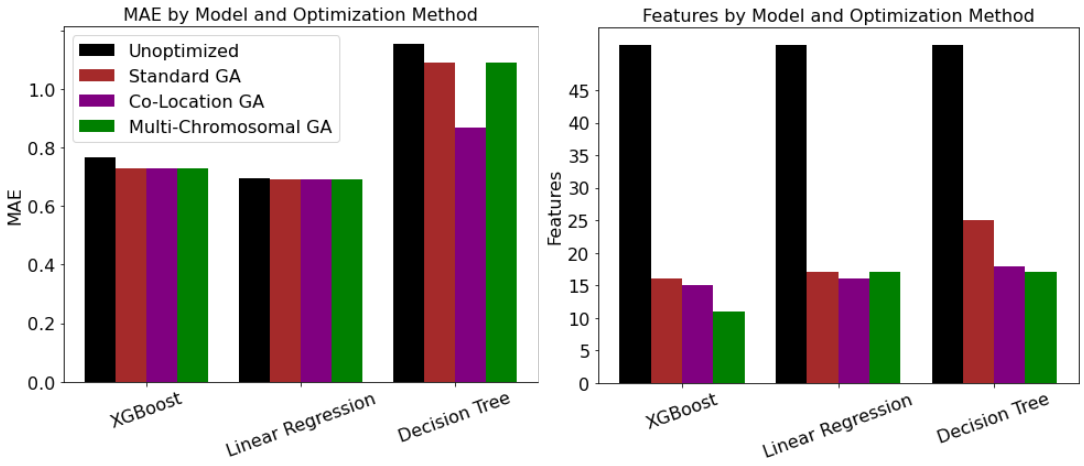


Figure 19: *Comparison of the best MAE achieved (left) and minimum features required to achieve optimal results (right) by each model, per optimisation method.*

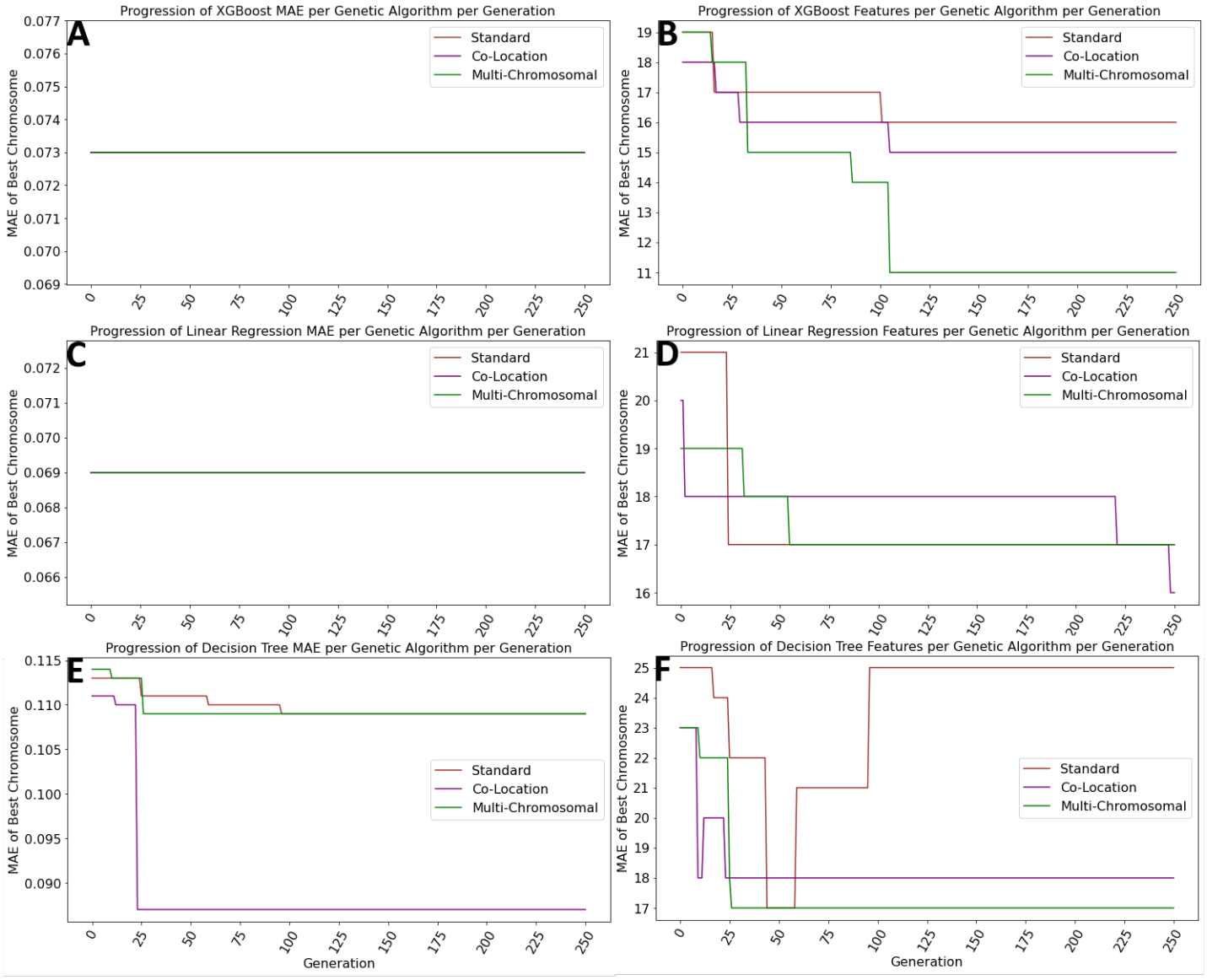


Figure 20: *Comparison of the evolution of MAE and minimum features required for the XGBoost, Linear Regression and Decision Tree models, for each genetic algorithm methodology.*

In terms of optimisation of the minimum features required, the CLGA and MCGA displayed superiority over the SGA, for the XGB (SGA: 16, CLGA :15, MCGA: 11) and DT models (SGA: 24, CLGA :16, MCGA: 15). However, CLGA reduced the feature requirement of LR, from 16 for SGA and MCGA to 15 (Figure 19, right).

No GA produced improvement of the MAE achieved by XGB and LR, these plateaued at MAEs of 0.073 and 0.069 respectively (Figure 20, A, C). DT MAE was improved by each GA methodology, achieving the lowest MAE after 25 generations of CLGA evolution. MCGA and SGA achieved equal MAE, but MCGA achieved this after 25 generations, whereas SGA achieved maximum performance after 100 generations (Figure 20 E).

Each GA methodology gradually reduced features required (Figure 20 B, D, F), except for SGA for DT optimisation, which reduced the features required from 25 to 17, but back to 25 when a superior MAE was achieved (Figure 20, F). MCGA achieved the lowest feature requirements for both the XGB and DT algorithms (Figure 20, B, F). CLGA achieved the lowest feature requirement for LR (Figure 20, D).

Evaluation:
This benchmarking analysis found no GA methodology to significantly improve the MAE achieved by the best performing XGB and LR models. This may be due to automatic feature selection built-in to these models, overriding the feature selection performed by the GA. This may indicate that few powerful variables are dominating and dictating predictions and the resulting MAE. Interestingly, the DT models were gradually improved by each GA, both in terms of MAE and the minimum features required. This may be due to the natural incorporation of more variables into the branching pattern of DT, and thereby a greater sensitivity to the variables included in tree construction.

The analysis of the progressive evolution of each model in terms of MAE and feature minimisation (Figure 20, A:F) demonstrated that the MCGA and CLGA models achieved the best MAE and feature requirements for each model, and faster than the SGA. MCGA appeared to achieve the lowest number of features required for each model faster than any other GA, except for when minimising the features required by LR (Figure 20 D); in this instance, CLGA achieved the lowest requirement on the second last generation, but MCGA achieved the second lowest feature requirement (17) 175 generations before CLGA did. This indicates that for optimization of more computationally expensive processes, MCGA may help achieve near-optimal results in a fraction of the time as CLGA, and even faster than an SGA.

Together, this implementation, results and evaluation of this benchmarking section comprised completion of Objective 8 (See Section 1.1, Table 1).

5.8 Discussion

In this study, 160,000 transactions were analysed and modelled by XGB, LR and DT regression models. These models were then optimized through GA-mediated feature selection. Specifically, the accuracy with which these models could detect errors in the Zillow Zestimate model, and while using the fewest numbers of features, was evaluated. Three alternative GA methodologies, including two novel GA, were trialled. The result of this analysis was the identification of two novel GA methodologies that outperformed the SGA both in terms of best MAE achieved, minimum features required, and in terms of speed of evolution (that is, they achieved optimal results with fewer generations and less evolution time required).

Despite these encouraging results, this study was limited by a number of factors. First, computational constraints prevented the analysis and optimisation of SVR. These models may have outperformed the alternatives, based on their ability to predict house prices in the literature. However, as these models are sensitive to highly dimensional data, they may never have been suitable candidates for this specific analysis. Computational constraints also limited the extent to which each evolutionary algorithm could act on each model. Performing thousands of generations of optimisation may have substantially improved results (as can be seen from the TSGA, which minimised the objective function by 68% after 100,000 generations). However, it should be noted that one purpose of conceiving novel GA was to compensate for computational costs associated with the SGA rubric. This objective was achieved through the CLGA and MCGA methodologies, which each achieved better prediction results, and required fewer features, while also achieving optimal results faster than the SGA. Therefore, this analysis was successful in its aim of proposing novel evolutionary algorithms.

This analysis also successfully delivered an exploratory analysis of the price prediction errors across Los Angeles. Although the findings that missing data in the air conditioner system field affects effective price prediction, and that there is greater prediction error with the size and number of rooms in a house, this exploratory analysis identified two actionable insights through which price prediction practices in Los Angeles can be improved. First, a crux of incorrect property price predictions in the Laguna and Huntington areas (Figure 8, circled red) indicates that this are likely displays unique characteristics that render it difficult to model. Further, the analysis of prediction error as a function of time (Figure 7) indicates that not only are there periods of the year when prediction errors are more frequent, but that there may be external factors acting on property prices, not accounted for in the dataset. That is, that there are periods of the year when more properties are purchased, which drives up prices through supply/demand dynamics.

As this analysis centred around the modelling and analysis of Zillow's private Zestimate model, it is difficult to compare the performance of this analysis to that of others in the literature. A search of the literature revealed no similar analyses of this particular dataset or model. However, the results achieved here fall between those achieved by the leaders of the private and public leader boards on the Kaggle competition page. In addition to the difficulty in comparing the performance of machine learning model predictions to similar results in the literature, it is not possible to compare the results of the CLGA and MCGA to others in the literature. As these GA are novel and this analysis is the first investigation of their utility, there are no examples in the literature to which these results may compared. However, given that the SGA is the industry standard, comparison of the CLGA and MCGA results to the SGA methodology is a valid benchmark, and a good indicator of the potential utility of these novel methods in additional domains.

Table 4 - Comparison of Performance

Source	Models	Optimisation Method	MAE	Features
dset/aichoo.ai	Unknown	Unknown	0.0632	Unknown
dset/aichoo.ai	Unknown	Unknown	0.0632	Unknown
This Analysis	Linear Regression	Co-Location Genetic Algorithm	0.069	15
This Analysis	XGBoost	Multi-Chromosomal Genetic Algorithm	0.073	11

6 Conclusion and Future Work

The question answered by this investigation was: *To what extent does combining additional biologically inspired GA enhancements with current best practices improve algorithmic efficiency, in a United States house price prediction application?* This investigation found that by incorporating the biological (genetic) phenomena of gene clustering (co-location) and the distribution of genetic material across multiple chromosomes, substantially improved GA optimization. Specifically, these novel algorithms results in models that performed better than those optimized by SGA, and also achieved optimal results faster than SGA. This finding is significant for any fields in which GA are deployed for optimization of feature selection and may have impacts in those experiments where GA are deployed for optimization of other objectives, such as hyperparameter tuning. This investigation also has implications for those optimization tasks that involve computationally expensive components; the finding that MCGA and CLGA algorithms achieve optimal results faster than standard methods means that fewer generations and fewer iterations may be required for optimization, meaning reduced computational cost.

Although this analysis focused on property price prediction, the analysis dataset was restricted to Los Angeles in the United States of America. Therefore, the insights garnered from this investigation may not apply to geographies with different property price dynamics.

Further, this analysis compared the performance of various GA methodologies, but only using a single dataset and application. To fully understand whether the novel methodologies presented in this analysis are superior to standard methods, experimentation using multiple datasets and data sources would be advised.

Computational constraints also limited the depth of exploration in this study. It was due to the computational cost and time required, that models that may have potentially outperformed LR and XGB, were not analysed here. Similarly, this computational constraint also limited the number of generations each GA could run for. The impact of this is emphasised by the finding in Figure 18 D, that the CLGA achieved a substantial improvement in the second last (249^{th}) generation.

Future analyses should aim to investigate the utility of these novel GA across a range of house price prediction datasets. Encouraging results in that exercise would then motivate investigations of datasets and data sources outside the domain of property price prediction, as in theory, these GA methodologies should be applicable to all feature selection problems, across domains.

Similarly, these GA should be applied to optimization of hyperparameter tuning. As this analysis found no GA to substantially improve predictions of the XGB and LR models through optimisation of feature selection, optimization of the hyperparameters may have yielded superior results.

Finally, and based off the findings of the exploratory analysis: an investigation into the unique characteristics of the Laguna and Huntington Beach areas of Los Angeles is motivated. A better understanding of that local property market may yield improved price predictions.

Acknowledgement

I would like to take this chance to thank those who helped and supported me through to completion of this investigation and technical report. First and foremost, Dr. Catherine Mulwa, who provided consistent and vital support and feedback throughout the course of

the investigation. I also appreciate the support of my friends, who showed great patience when Genetic Algorithms were the only thing on my mind. Although the struggle to complete this investigation could not compare to the struggles many of my colleagues and friends endured through the pandemic, I am grateful for the opportunity to contribute to the field.

References

Blanquero, R., Carrizosa, E., Jiménez-Cordero, A. and Martín-Barragán, B. (2019). Functional-bandwidth kernel for Support Vector Machine with Functional Data: An alternating optimization algorithm, *European Journal of Operational Research* **275**(1): 195–207.

Bzdok, D., Krzywinski, M. and Altman, N. (2018). Points of significance: Machine learning: Supervised methods.

Dong, S., Wang, Y., Gu, Y., Shao, S., Liu, H., Wu, S. and Li, M. (2020). Predicting the turning points of housing prices by combining the financial model with genetic algorithm, *PLoS ONE* **15**(4): e0232478.
URL: *https://doi.org/10.1371/journal.pone.0232478*

Du, Q., Wu, C., Ye, X., Ren, F. and Lin, Y. (2018). Evaluating the Effects of Landscape on Housing Prices in Urban China, *Tijdschrift voor Economische en Sociale Geografie* **109**(4): 525–541.

Gündüz, H. (2019). Comparison of different dimensionality reduction methods in the detection of Parkinson's disease, *Eur. J. Sci. Technol.* (17): 1164–1172.

Hambali, M. A., Oladele, T. O. and Adewole, K. S. (2020). Microarray cancer feature selection: Review, challenges and research directions, *International Journal of Cognitive Computing in Engineering* **1**: 78–97.

Hussain, A., Muhammad, Y. S., Nauman Sajid, M., Hussain, I., Mohamd Shoukry, A. and Gani, S. (2017). Genetic Algorithm for Traveling Salesman Problem with Modified Cycle Crossover Operator, *Comput. Intell. Neurosci.* **2017**.

Jha, K. and Saha, S. (2021). Incorporation of multimodal multiobjective optimization in designing a filter based feature selection technique, *Applied Soft Computing* **98**: 106823.

Kaggle (2018). Zillow Prize: Zillow's Home Value Prediction.
URL: *https://www.kaggle.com/c/zillow-prize-1/data?select=properties_2017.csv*

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F. and Ratti, C. (2020). Understanding house price appreciation using multi-source big geo-data and machine learning, *Land Use Policy* p. 104919.
URL: *http://www.sciencedirect.com/science/article/pii/S0264837719316746*

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. and Liu, H. (2017). Feature selection: A data perspective.
URL: *https://doi.org/10.1145/3136625*

Liu, R. and Liu, L. (2019). Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm, *Soft Computing* **23**(22): 11829–11838.

Muñoz-Romero, S., Gorostiaga, A., Soguero-Ruiz, C., Mora-Jiménez, I. and Rojo-Álvarez, J. L. (2020). Informative variable identifier: Expanding interpretability in feature selection, *Pattern Recognition* **98**: 107077.

Ng, A. (2015). Machine Learning for a London Housing Price Prediction Mobile Application, *Technical report*.

Owusu-Ansah, A. (2013). A review of hedonic pricing models in housing research, *A Compendium of International Real Estate and Construction Issues* **1**: 17–38.

Saeid, M. M., Nossair, Z. B. and Saleh, M. A. (2020). A microarray cancer classification technique based on discrete wavelet transform for data reduction and genetic algorithm for feature selection, *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, Institute of Electrical and Electronics Engineers Inc., pp. 857–861.

Su, T., Li, H. and An, Y. (2021). A BIM and machine learning integration framework for automated property valuation, *Journal of Building Engineering* **44**: 102636.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S2352710221004940*

Truong, Q., Nguyen, M., Dang, H. and Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques, *Procedia Computer Science* **174**: 433–442.
URL: *http://www.sciencedirect.com/science/article/pii/S1877050920316318*

Vlašić, I., urasević, M. and Jakobović, D. (2019). Improving genetic algorithm performance by population initialisation with dispatching rules, *Comput. Ind. Eng.* **137**: 106030.

Wang, D., Zhang, Z., Bai, R. and Mao, Y. (2018). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring, *J. Comput. Appl. Math.* **329**: 307–321.

Xu, J., Pei, L. and Zhu, R.-z. (2018). Application of a Genetic Algorithm with Random Crossover and Dynamic Mutation on the Travelling Salesman Problem, *Procedia Comput. Sci.* **131**: 937–945.

Xu, Z., Zhang, T., Keung, J., Yan, M., Luo, X., Zhang, X., Xu, L. and Tang, Y. (2021). Feature selection and embedding based cross project framework for identifying crashing fault residence, *Information and Software Technology* **131**: 106452.

Xue, B., Zhang, M. and Browne, W. N. (2015). A comprehensive comparison on evolutionary feature selection approaches to classification, *International Journal of Computational Intelligence and Applications* **14**(2).

Zillow_Research (2020). 2020 Was a Surprisingly Strong Year for Housing. Here's Why 2021 Will be Stronger. - Zillow Research.
URL: *https://www.zillow.com/research/november-2021-sales-forecast-28499/*