National College of Ireland

Housing Price Prediction and Classification Based on Crime Occurrence using Machine Learning Algorithms: Ireland

Salam Adekunle Adedokun
x18156037

School of Computing
National College of Ireland

MSc Data Analytics – 2019/2020

Supervisor: Dr. Catherine Mulwa

Housing Price Prediction and Classification Based on Crime Occurrence using Machine Learning Algorithms: Ireland

Salam Adekunle Adedokun
x18156037

January 28, 2020

Abstract

Housing is a basic need for humans and the acquisition of a house is an action that should be taken carefully because of the financial implications. This project aims to examine the comparative performance of both predictive and classification models and how the crime occurrence, distance to both the nearest primary school and bus-stop can improve model performance. The crime occurrence variables include attempts to murder, assaults, burglary, theft, fraud, drugs, weapon offences, damage to property and social code offence. This model would help real estate investors and prospective house buyers to predict and classify housing prices thereby reducing loss for real estate developers and giving the negotiating power to prospective house buyers. Comparative analysis performed on the data mining techniques; generalised linear model, ridge, lasso, support vector machine and random forest was evaluated based on their Mean Absolute Error and Root Mean Square Error while the comparative analysis of classification data mining techniques: random forest, C5.0, k-nearest neighbours, support vector machine and multinomial logistics regression was evaluated based on accuracy. The output of the algorithms was visualised with tableau to give a clear insight on how the models performed for both the prospective house buyers and real estate investors.

Area Machine Learning: is the capability of a computer to learn by finding consistent or hidden patterns in a data to enable it accurately classify or make predictions, and this is used in this project for house price prediction and classification.

1 Introduction

Shelter is a major biological requirement for human survival, and according to Maslow (1943) it is one of the fundamental psychological needs for humans to function properly. A house serves as a shelter for a person, group of people or family. Hence, the need for a household and its environs to be safe. The security level in an area is a measure of safety and has a direct relationship with both the cost of living and housing price in that area. The acquisition of a house is a lifetime goal for most prospective house owners and usually involves a huge fraction of their net worth or savings, in the United States of America (USA) studies show that there has been prevalent sales of overvalued houses which is caused by lack of sufficient and reliable information from marketers and dicey advertisements. The project aims to show the importance of safety and distance to schools when valuating a house, by creating a model for the prediction of housing prices to foster a more accurate decision for real estate investors, tax assessors, house developers, prospective house owners, insurers and mortgage lenders (Frew & Jud 2003). Furthermore, this project helps bridge the information gap between the Realtor's and the prospective house buyers, helping buyers to have an estimate of the property, also gives a more accurate housing price prediction model, In order for investors to make great decisions based on the current value of a house and future value. Since it would be inappropriate for the price of a house to be fully determined by the house owner, the determination of the price of a house is derived from various factors like building structure, building size, location, crime rate in the area and distance to the nearby schools. The various types of crime rate considered for the purpose of this project are; attempts to murder, assaults, burglary, theft, fraud, drug offence, weapon offences,damage to property, social code offence and Offences against government

1.1 Project Background and Motivation

The economy of a country is now one of the relevant factors to a society more than ever in history. In 2008, the same period when there was a banking crisis in Ireland, after the intervention of the government to provide liabilities, the was a growth in the Irish housing bubble which was cause d by a strong belief of the Irish to store wealth through Property according to Regling & Watson (2010). this bubble was in existence from the 1990s to the late 2000s. This led to a sharp drift from a competitive economy to a credit driven economy due to the wide acceptance of most inexperienced individuals who believed that accessing cheap credit could make them real estate professionals Kelly (2009), this action had a huge impact on the banks by increasing their lending percentage from 60% in 1997 to 200% in the late 2000s. The current economy shows a striking resemblance to the USA lending economy.

In 2007, the United States experienced a great recession which had a significant impact of housing development , this was measured through the decline in the number of building permits issued, both the start and completion of housing construction reduced, and the total housing sales experienced a downward slope in the housing sector that had a significant effect on the Gross Domestic Product (GDP) in 2007. The crash of the housing market caused more financial damage to house developers and real estate developers who have major investments and had a serious challenge of making their initial investment. The ability to be able to predict the price of a house would be of tremendous advantage to investors because, this would help them make use of present market condition in order

to evaluate the future market and implement proactive actions. The motivation of this project is to show how the safety of an area where the house is located and the distance of the house to nearby schools would help improve the accuracy of house prices prediction.

According to Baumann & Friehe (2013) crime is identified as a social phenomenon, and the panic or scare it causes to a person, family and various economic activities makes it one of the major social deficiencies. Apart from the loss of monetary value items that crime causes, crime also has indirect implications to the area which includes increased insecurity and anxiety. The consequences of crime are experienced distinctively in various areas but most neighbourhoods experienced violence in an event of crime. The research carried out in Spain by Buonanno et al. (2013) showed how crime cost the economy an annual loss of about a trillion dollars. After considering the impact of crime, it is also important to understand the common causes of crime in an area like areas with little or no social amenities, bad road network, minimal socializing and commercial zones (Bars, Restaurant, Banking facilities etc) have a higher chance of experiencing crime. It is important to note that areas with low crime rate would attract new and larger investments, more social amenities, these developments would increase the value of the houses in the area.

Another important factor during the decision making for most home buyers is the distance of the house to nearby schools. Hence, there is usually a high demand for houses closer to schools which translates to a higher price. Previous works placed more importance on the effects of school quality, but it is also necessary to consider the closeness to the nearby schools. This is because families prefer to live closer to schools for transportation and convenience, Also, a more educated area would help shape the mentality of residents in that area and boost the areas economic growth and translates to a reduction in crime rate. Crime rate and violence in an area usually has a ripple effect on the influx and out-flux of residents of an area, where increased crime rate leads to higher out-flux and lower influx of residents and vice-versa for reduced crime rate. The ability to accurately predict housing prices is an invaluable asset to property investors because, it helps investors to create an investment portfolio that mitigates risk of financial loss and optimize maximum investment profit. The strong correlation between the economy and the housing price is also a motivating factor for developing a model that predicts house prices (Pow, Janulewicz, Liu, 2014).

This project is beneficial to prospective house buyers, house developers and investors for the proper valuation of houses, to help make guided decisions on investment, also, it would help sellers to eliminate the cost of auctioning with buyers being comfortable with the bid price knowing it is not an outrageous cost due to the accurate model (Mukhlishin et al. 2017). The innovation of this project is the inclusion of crime occurrence to build a housing price appraisal model, the justification for this addition is to emphasize the importance attached to safety when acquiring a house. For the implementation of this project about ten machine learning algorithms were implemented to both classify and predict housing price and to determine the most accurate algorithm for this model feature engineering was performed on the merged dataset and important variables were selected to improve the performance of the models.

1.2 Research Question

Prediction of housing prices is significant because it helps get a good bargain during house acquisition, this project would predict and also classify housing prices based on the

economic importance of safety and education and further more improve the accuracy of housing prediction. This project aims to answer the following research questions,

RQ: *"To what extent can machine learning algorithms be used for prediction (generalised Linear Model, ridge regression, lasso regression, support vector machine and random forest) and classification (random forest, C5.0, k-nearest neighbours, support vector machine and multinomial logistics regression) of housing price, based on crime occurrence (attempts to murder, assaults, burglary, theft, fraud, drug offence, weapon offences,damage to property, social code offence and Offences against government) reported at all garda stations, distance of houses to the closest bus-stop and primary school in Ireland?"*

1.3 Research Objectives and Contributions

The first objective was to take a critical investigation into the housing industry, price prediction and related works from 2002 to 2018. Other objectives required to be accomplished for the research questions to be answered are outlined on Table 1 Below.

Table 1: Research Objectives

Index	Description	Evaluation Method
Ob 2	Feature Engineering	
Ob 3	Implementation, Evaluation and Results of Housing Price Prediction Models	
Ob 3(a)	Implementation, Evaluation and Results of Support Vector Regression.	RMSE and MAE
Ob 3(b)	Implementation, Evaluation and Results of Random forest.	RMSE and MAE
Ob 3(c)	Implementation, Evaluation and Results of Lasso Regression.	RMSE and MAE
Ob 3(d)	Implementation, Evaluation and Results of Ridge Regression.	RMSE and MAE
Ob 3(e)	Implementation, Evaluation and Results of Generalised Linear Model Regression.	RMSE and MAE
Ob 4	Implementation, Evaluation and Results of Housing Price Classification Models	
Ob 4(a)	Implementation, Evaluation and Results of Multi-nominal Logistics Regression	Accuracy
Ob 4(b)	Implementation, Evaluation and Results of Support Vector Machine.	Accuracy
Ob 4(c)	Implementation, Evaluation and Results of Random Forest.	Accuracy
Ob 4(d)	Implementation, Evaluation and Results of C5.0	Accuracy
Ob 4(e)	Implementation, Evaluation and Results of K-Nearest Neighbors	Accuracy
Ob 5(a)	Comparison of Implemented Models	
Ob 5(b)	Comparison of Implemented Models Against Existing Models	

This project would contribute to the success of real estate companies by aiding their decision on areas to invest, based on possible crime rate and this would enhance the chances of investment profitability. The Classification of the housing prices would enable investors understand the economic bracket of the house and gives more insight on the value of the house. Also, this project would help prospective house owners negotiate a valuable deal for their house, while taking into consideration the value of the house and safety of the geographical zone where the house is located.

The remaining sections of this project are as follows: Chapter 2 contains the critical review of related works on housing prediction models, Chapter 3, shows the methodology, design specification and feature engineering of the project, Chapter 4 gives detailed illustration of the implementation, evaluation, results, software and technology used for the implementation of the project. Chapter 5 contains the discussion and comparison of implemented models. Lastly, the conclusion and future works for possible improvement of the project.

2 Literature Review of Housing Price Prediction

In this section, the previous works on prediction of housing prices are constructively criticized based on the techniques, datasets, methodology, features and results. The scope of this review is within a span of 16 years (2002 - 2018), within the span of this investigation, to the best of the candidate's knowledge, the review below supports the novelty of this project.

2.1 The Irish Housing Crisis

The housing crisis in Ireland has evolved strongly over the last two decades and these events have been classified into three phase: the first phase being the Celtic Tiger Era that started in the early 1990s, in this era there was an evident rise in the economic indicators of the country that led to economic growth and attracting more immigrants into the country, according to *Vital Statistics* (2014) while the population in Ireland increased by 20% the household also increased by 43%, regardless of this expansion, the demand of housing also rose drastically. this demand led to a rise of about 420% in the price for a new house in 2007 compared to the price in 1991 and this inn turn doubled the mortgage. Larragy (2014) stated that the rise in the debt of the country due to the increased value of homes has a good effect on the economic growth. The second Phase in the economic crisis was when the economy slowed down from 2007 to 2012 where from the peak the price of houses fell by about 60% and was recorded as one of the most extreme house collapse, this economy crash affected all sectors including the banking industry. The third and most resent phase of the housing crisis is often regarded to as the re-balancing due to the gradual and rapid increase in housing prices, the phase is a continuation of the rise after the fall in the second phase, due to insecurity of income and rising housing rent, most families are dependent on welfare in this phase.

2.2 A Critique of Techniques, Method, Features and Evaluation of Housing Price Prediction

In the past years the approach practiced in an attempt to predict housing prices was the machine learning and regression method Park & Bae (2015), various amount of factors are kept in consideration like house age, inflation rate, mortgage rate, Net migration and government policies. Shinde & Gawande (2018) also considered independent variables like size of the house, house age, location, size of garage and the overall quality of the house, these features were evaluated for about 15,000 houses, the following models were implemented; Support vector machine, decision tree and lasso regression, to measure their performance , these models were evaluated using Root Mean Squared Value (RMSE), Mean Squared Value (MSE) and R-squared value. The result of the evaluation shows that the decision tree outperformed the other models while the lasso regression had the lowest performance. An innovative research was performed by Mukhlishin et al. (2017) where they used quite distinct features like tax on land , asset price, location of the house and its physical condition, the techniques used in this project were k-nearest neighbours, fuzzy logic and artificial neural network, after evaluation using mean absolute percentage error (MAPE) , the fuzzy logic had the most accuracy compared to the other two models. Varma et al. (2018) applied the forrst regression, linear regression and neural networks to their dataset containing supermarket locations, railway station, parks and hospitals.

Park & Bae (2015) used public schools ratings, mortgage rate as their features while performing modelling with Ripper, C4.5, and naïve Bayes, after evaluation Ripper was the best performing model. Chiarazzo et al. (2014) tried to find patterns in the features of the amount of pollution damage caused, number of industries in that area by implementing an ANN model

Azadeh et al. (2012) had a critical look at how the economic factors like seasonal income, inflation rate, demand of housing, supply of housing and the relative investment in that location with use of data mining techniques like fuzzy cognitive map and fuzzy regression both had no difference in variance and this could pose as an obstacle to building a model with them

2.2.1 A Review of Regression and Classification Models

When dealing with linear data regression models have proven to be great performers, but when some of the features are non- linear the model would begin to encounter troubles, these troubles have been put to an end by a work of Bin (2004), with his discovery that finds a way to gather data from the geographic information system and keeps all of the data in a locational attribute. After a wide comparison of other ways to handle the locational data, this project used the same approach to enable the models parse and find insight in the datasets. Kahveci & Sabaj (2017) has strongly made a claim that the distance to schools is an important factor used in the prediction of housing prices, with the experiments the y carried using ANN and this was possible because of the ability od ANN to work with non-linear models.

The classification of housing prices is an area of knowledge that has only been exploited by a few researchers. This approach would help us explore unique techniques that could have been restricted to only regression problems and rage of classification gives better understanding to the end users at the presentation tier (Yu & Wu 2016).

2.3 An Investigation on the Effect of Crime Rate and Schools on Housing Price

It was once stated by Cohen (2008) that the feeling as a whole that you are safe in an area has more impact on residents than the statistics of crime in that area. The effect of crime can have different lanting effects on people as some would tend to experience fright, anxiety, and not just loss of their physical properties, when residents of an area begin to feel this way about their environment, they begin to emigrate Gibbons (2004), which would lead to a drop in demand for houses in that area and the price would consequently drop. In the early 1900s before the drastic crime drop In USA, it was bbelieved that there is a correlation between socioeconomic factors in an area and the crime rate, but this was debunked by Cook (2008) who performed experiment in over 200 counties of the USA, the socioeconomic factor does not have any correlation with crime

When considering the factors that affect the housing price of a place the most, it is assumed that schools are the most dominant factors for predicting housing price, Nguyen-Hoang & Yinger (2011) worked on the impact of the quality of school on the housing prices but also the distance to school and the distance to the nearest busstop is one of the frequent questions asked by home buyers. And most houses located to schools are always averagely higher than the other houses in that are, because of the economic benefits attached to the value that a nearby school gives. The previous works carried out

by Rosiers et al. (2002) and Espey et al. (2007) in an attempt to measure the distance to the closest school failed because their measurement was on a range basis and no actual values were recorded. Metz (2015) also performed a similar study using quite vague form of measurements that were non in a continuous measurement form but he was still able to make a deduction that nearby houses to schools had more economic benefits.

2.4 Identified gaps

From the above literature review one of the works stated how security could be a determinant of price in an area, hence we used the safety metric in conjunction to the two variables that were considered the most important in past works, these are distance of the house to school and distance of a house to the busstop. A summary of previous works, showing their respective dataset size, data source, data mining techniques and features is shown below in Table 2.

Table 2: Comparative study of Previous works on Housing Price Prediction

Features	Sample Size	Techniques	Dataset	Evaluation Metic	Result	Author & Year
Overall condition of the house, Location, Year house was built, Numbers of Bedrooms and bathrooms, Garage area and number of cars, swimming pool area, selling year and Price house is sold.	15000	Logistic Regression, SVM, Lasso Regression and Decision Tree	Housing prices from Kaggle	R-squared	84.64%	Neelam, Kiran Gawande (2018)
Building types,Land area, Age, Building height, Number of bedrooms, Number of living halls, Number of bathrooms, Parking lot dummy, Locational attributes, Distance to riverbank, Subway station, University, Hypermarket, Department store, Supermarket, Night market, Hospital, Police stations, Fire station	3991	Multiple regression & SVMmodel	Taipei city's properties (2007 and 2010)	R-squared	64%	Jieh-Haur CHEN (2017)
Value of sales, value of taxable object land (NJOP-L), sales value on taxable object building, house age, house condition and land strategic location	7500	Fuzzy logic artificial neural networks, K-Nearest Neighbors algorithm	Indonesia housing Dataset	MAPE	88%	M. F. Mukhlishin, et al. (2017)
Based on location, house type, size, build year, local amenities,	1300	Ridge, Lasso and Gradient boosting	Ames, Iowa	RMSE	0.1126	Sifei Lu, Zengxiang Li (2017)
House prices, numbers of bathrooms, bedrooms, living rooms, Agricultural, Residential High Density, Residential Low Density, Residential Low Density Park	1460	Lasso, Ridge, SVM regression, and Random Forest regression and classification methods include Naive Bayes, logistic regression, SVM classification, and Random Forest classification.	Ames, Iowa	RMSE and Acuraccy	Accuracy: 0.6913 and RMSE: 0.5269	Yu and Jiafu Wu (2016)
Real estate, public school ratings, and mortgage rate data	15,135	C4.5, RIPPER, Naïve Bayesian, and AdaBoost	Virginia housing data	RMSE	0.201	B. Park and J. K. Bae (2015)
Land use attributes of the collected properties.	193	Artificial Neural Networks (ANN)	On-line Dataset of real estate database platforms (October 2012)	R-value	0.83	Vincenza Chiarazzoa (2014)

Conclusively, this project looks at the comparative analysis of regression models (Generalised Linear Model, Ridge regression, lasso regression, support vector machine and random forest) and classification models (random forest, C.50, k-nearest neighbours, support vector machine and multinomial logistics regression) with the aid of evaluation metrics (RMSE, MAE and Accuracy) to understand the impact of crime occurrence, distance to primary school and distance to busstop. The dataset for this research was sourced from the Ireland Property Price Register database and Central Statistics Office Ireland, Transport for Ireland in accordance to GDPR.

3 Housing Price Prediction Methodology and Design Specification

3.1 Methodology

The Housing Price Prediction Methodology used for this project was been modified to suit the goal of the project. The addition of feature extraction to the methodology helps determine the variables with the highest contributing factor in the prediction of house prices. The sequential process that the Housing Price Prediction Methodology entails would be discussed below, and it is illustrated in Figure 1.

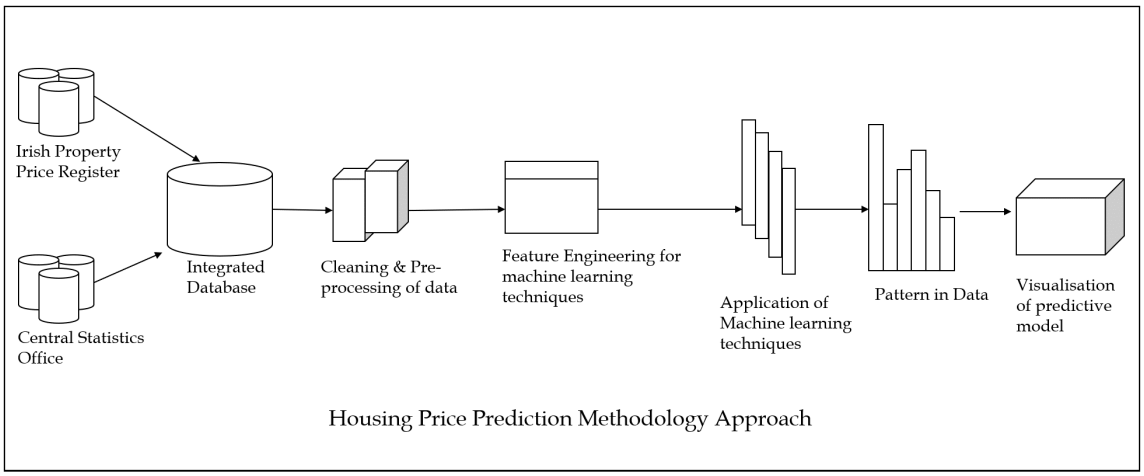

Figure 1: Housing Price Prediction Methodology

Integrated Database: This stage is also known as the data compilation stage and it is the initial process of the Housing Price Prediction Methodology. For this project Four datasets from four distinct databases were used. The major dataset which contains the house prices sourced from The Irish Property price Register database, The second dataset is the bus-stop location dataset, this data set was downloaded from the Transport For Ireland Repository, the third dataset consists of all the primary schools this dataset was retrieved from the Maynooth university Repository and the final dataset was sourced from the Central Statistics Office, Ireland, this dataset contains recorded crime in Ireland.

Pre-processing and Cleaning of Data: The datasets in the database are individually cleaned by removing or imputing missing values, converting some of the variables to factors where such cases rise, then based on the key columns the datasets are transformed with the aid.

Feature Engineering for the Data Mining Techniques: An Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA) was performed on the merged dataset to get insight on the dataset, the Normality, Correlation, outliers and inter-quartile range. The insights gotten from the EDA helps validate the usability of the dataset and helps make decisions on the elimination of variables with low correlation with the dependent variable. After validation of the dataset, the dataset is split into a Train set and Test set. After training of each model with the Training set, the Test set is used to evaluate the performance of the models. Before the implementation of the data mining techniques, based on their individual caveats, the datasets were scaled in order to create a standardized range of values for the algorithm.

Implementation of Data Mining Techniques: After the conclusion of feature Engineering, the train set is used to train the various algorithms and their accuracy is measured using the Test set. The accuracy evaluation methods used are RMSE and MAE.

Visualization and selection of the best predictive model: Various models find consistency and hidden patterns in the Train set and try to predict the dependent variable of the Test set. The accuracy of each algorithm is visualized to show the best performing data mining technique and their extent of accuracy to help the stakeholders decide the best algorithm for the model.

3.2 Design Specification

The primary aim of this project is to prove the hypothesis that the rate of crime, distance to nearest primary school and nearest bus-stop can improve the prediction of house prices. This chapter aims to explain the design process to be followed for the modelling of this project. The project does nor require the creation of a database for generating new data. Hence, the 2—Tier design option would be adopted which includes the client Tier (Tier 1) and Business Logic Tier (Tier 2). The modelling involves Data Sourcing, Geocoding, Data pre-processing, Feature Engineering, Implementation, Evaluation and Results comparison through Visualization. The design architecture of the house prices prediction model is shown below in Figure 2.

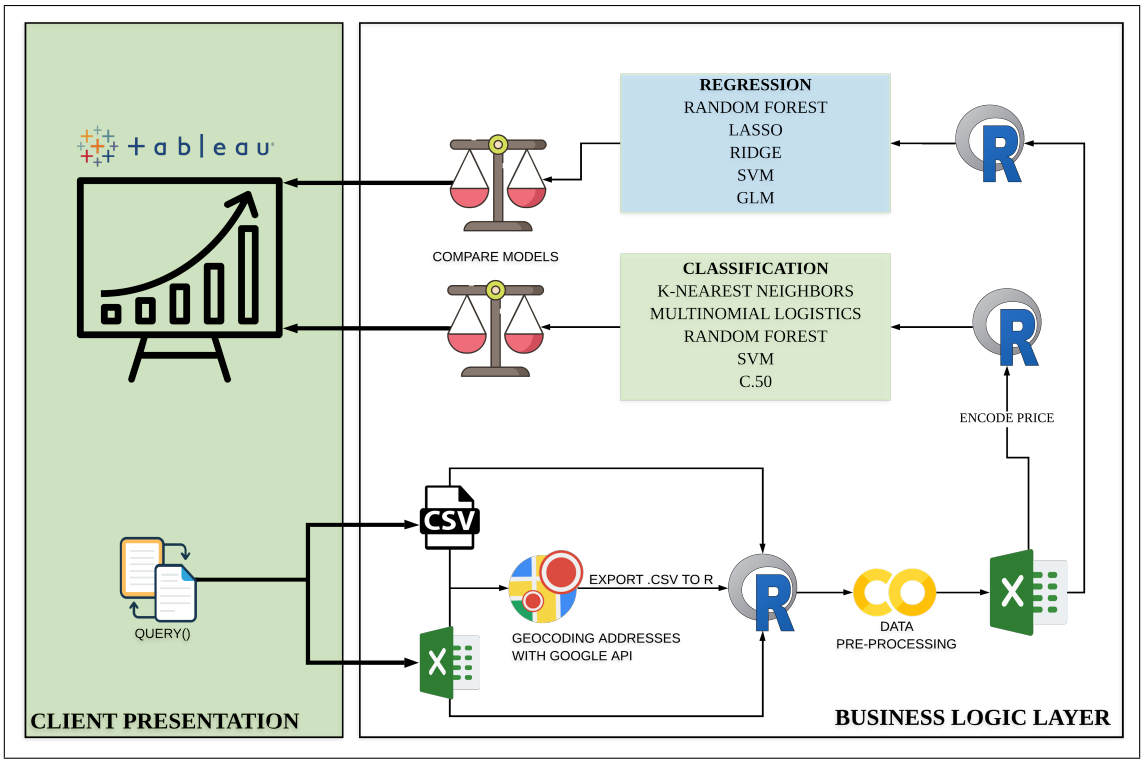

Figure 2: Housing Price Prediction Design

3.2.1 Data Source:

For the purpose of this project four datasets were used and were sourced from various trusted databases, this was done ethically. The datasets are outlined below;

- The dataset containing the price, address, sale date and property description of houses in Ireland (2012 – 2017) was sourced from The Irish Property price Register database.
- The dataset of all Bus-stops in Ireland and their location was sourced from Transport For Ireland data Repository.
- Complete dataset of the primary schools in Ireland and their address was downloaded from the Maynooth university Repository
- The garda recorded crime in Ireland (2003 – 2018) was sourced from the Central Statistics Office, Ireland.

3.2.2 Exploratory Data Analysis:

We explore our data set in the raw form to understand the dataset, find outliers and normalization and missing figures. Figure 3 shows that the price in its original form is positively skewed and to normalize the distribution the price variable was scaled with the logarithm function to result in a fairly normalised distribution shown below in Figure 4

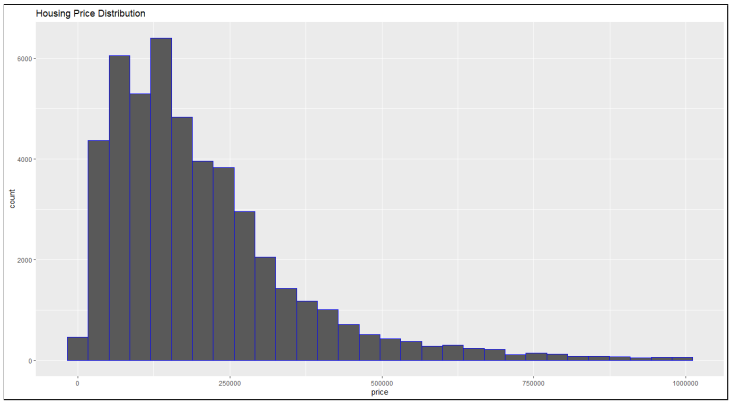

Figure 3: Original Price Distribution

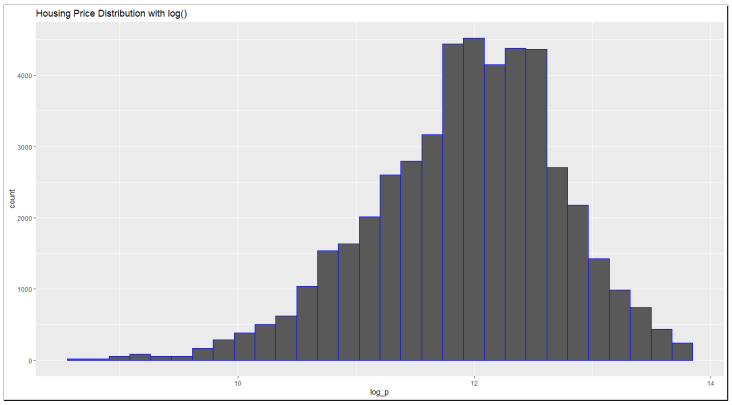

Figure 4: Scaled Price Distribution

The geographical exploration of the location coordinates is used to confirm the the generated coordinates by the Google Goecoding API, Figure 6 shows the locations that were wrongly geocoded and illustrated regions outside Ireland and totally away from Europe, these locations were identified and the row was deleted, also, a deeper look at the locations that illustrated outside the borders of Ireland in Figure 5 and some locations illustrated coastal zones, these locations were detected and deleted from the housing price dataset.

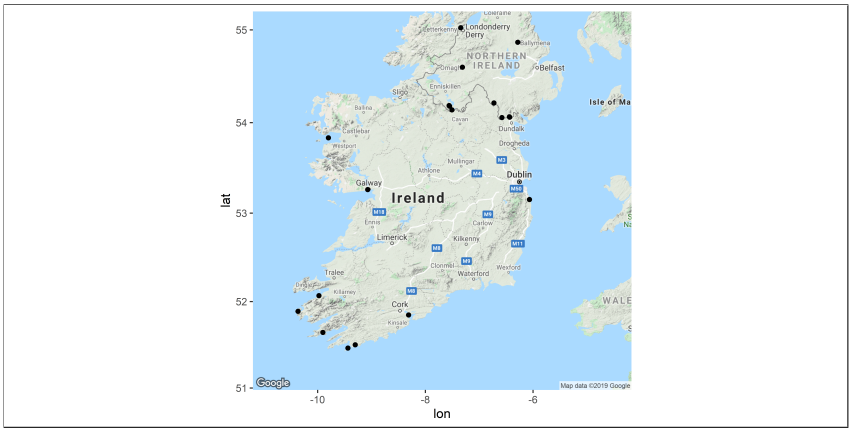

Figure 5: Map showing locations at the boarder and coastal areas

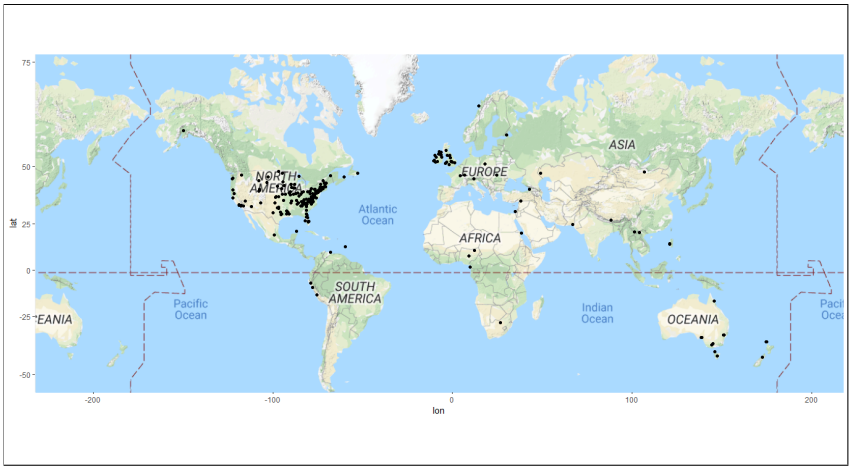

Figure 6: Map showing wrongly geocoded locations

Figure 7 and Figure 8 Shows how there has been successive growth in the housing price over the last 5 years, Though Dublin has always had a high price compared to other counties, the nationwide median housing selling price has grown steadily

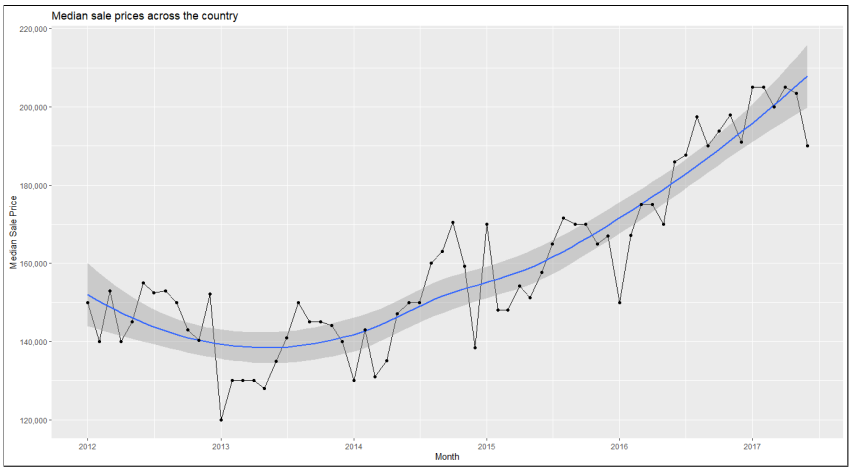

Figure 7: Ireland Median Housing Selling Price (2012-2017)

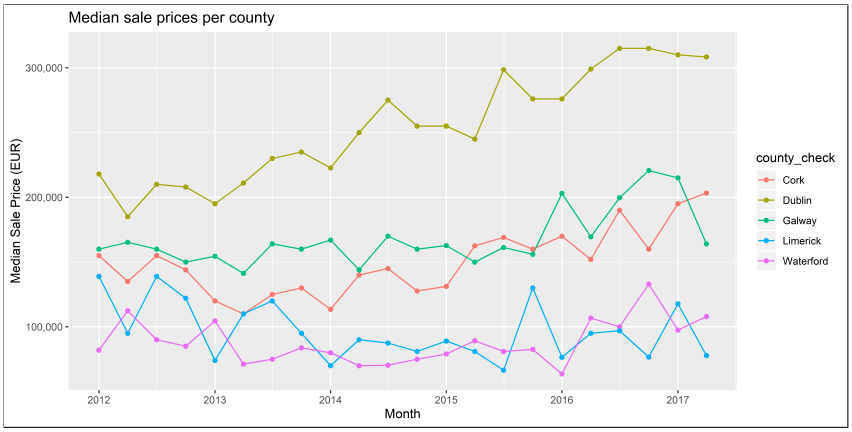

Figure 8: Median Selling Price Per County(2012-2017)

3.2.3 Data Pre-Processing

Data Cleaning: The downloaded datasets are cleaned individually using the R studio software, by checking for missing values and omitting "N/A" values because most of the missing values were in address column and could not be assumed, converting the date columns to a date class, removing irregularities in the address, removal of non-English housing descriptions, checking and removal of outliers in the house price dataset.

For the Classification algorithms we encoded the prices into three groups, the first is the "Low Cost" which represents houses that cost less than EUR120,000, the second group is called the "Medium Cost" it represents houses that cost more than EUR120,000 and less than EUR250,000 and houses that cost greater than EUR250,000 were classified as "High cost". Due to inconsistencies in all of the data sets, we used the datasets for the year 2015 only. After cleaning each dataset, they are saved individually in a (.csv) format.

Also, Houses with "No_Full_market_Price" were deleted, because this meant, the actual prices were not reported, hence it was not needed for the purpose of this project.

Transformation: After irrelevant features and missing values had been removed, and Boolean features had been changed to factors, we selected numeric features and normalized the data, this would allow the dataset to be modelled by some of the data mining models that require a normalized dataset. Also due to the nature of the housing price distribution, the price feature was passed through a logarithmic function to normalize the

skew, this step also helped improve the accuracy of the models. For the classification of the Price Range, the classification would make it easier for the model and thereby reduce the error rate that was noticed during the regression analysis.

3.2.4 Feature Engineering

This is the art of making more sense from a dataset by extracting the most relevant information from the dataset, for this project, we extracted distance to of schools and bus-stop to each house and created crime zones.

Geocoding of Addresses: Amongst the four datasets intended to be used for this project, only the Housing prices and the crime datasets lacked location coordinates. Hence, we used the address of each house to determine the coordinates, this was possible with the use of Google Geocoding API. With the aid of R-Studio the address of the houses were parsed into the Geocoding API which produces an output of two additional features to the table, these features represent the longitude and latitude of the house. While for the crime dataset, since we had only the station names and not the complete station address, we scraped the coordinates of each station from the garda station directory with a data scraping tool called Data-miner.

Distance Measurement: To determine the distance of a house to the closest school and bus-stop, we repeat the same process. For calculating the shortest distance between two geographic locations based on their coordinates, with the aid of R-studio each pair of coordinates are converted to a spatial dataframe, in this format the customized function below in Figure 9 measures the distance between a house and all the primary schools and keeps the shortest distance for each house. This same process is repeated for the bus-stops and the values would then be saved as a new features. These new created columns are added to the Property Price dataset. It is also important to note that one of the challenges in this stage is that Rstudio had to iterate the function in Figure 9 several times, hence, it demands a large memory for execution, this section on the code was implemented on the Google Colaboratory Notebook for R with larger resources.

```
nearest_p_school = apply(spDists(sp.Property_p, sp.primary_s), 1, FUN=min)
Property_p$nearest_p_school = nearest_bus_p_school
```

Figure 9: Function for shortest distance between two geographic locations

Zoning Crime Rate: For this phase, the assumption made was that most crime incidents are reported to the nearest police station of the incident occurrence. Hence, we attached each house to the nearest garda station along side the reported crime of the station. To achieve this, we find the index of the shortest distance between the houses and the garda stations based on their coordinates, with the aid of R-studio when each pair of coordinates are converted to a spatial dataframe, in this format the customized function finds the index of the shortest distance between a house and each station, with this index, name of the closest garda station and the total crime occurrence in that station is identified by the function in Figure 10. These new created columns are added to the Property Price dataset.

```
#To find the index of the minimum distance
try <- apply(nearest_station, 1, function(x) max(which(x == min(x, na.rm = TRUE))))

#using the index to generate nearest station name alongside its crime occurence
near_p_station = Crime[try, 2] #Name of the nearest station
Crim_oc = Crime[try, 18] #Total Crime  occurence in the station
```

Figure 10: Function for zoning crime based on crime occurrence

3.2.5 Description Of dataset

The merged dataset to be used for the implementation of this project has 49,912 rows and the attributes are described in Table 3 below:

Table 3: Features of Housing Price Dataset

S/N	Variables	Data Description
1	Sale Date	Date
2	County	Character
3	Sale Price	Integer
4	Full Market Price	Boolean: Yes/No
5	Vat Inclusive	Boolean: Yes/No
6	Address	Character
7	Latitude	Real
8	Longitude	Real
9	Electoral District	Character
10	Distance to nearest Bus Stop	Numeric
11	Distance to nearest Primary School	Numeric
12	Name of nearest Police Station	Character
13	Number of Crime Occurrence	Numeric
14	Price Range	Range: Low Cost =1, Medium Cost = 2, High Cost = 3

3.2.6 Evaluation and visualization:

The five regression machine learning algorithms implemented was evaluated by comparing their Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) while the other five classification machine learning algorithms was evaluated based on accuracy. The results of the models were compared are then visualized through tableau.

3.3 Conclusion

The methodology was adhered to throughout the implementation of the project and this helped to gain insight on the predictive power of each independent variable while taking into consideration which of the data mining techniques could accurately learn the model. During this project, the following tools/software's were used; Microsoft Excel, Tableau, Data Miner, Power BI, Google Colaboratory Notebook and RStudio.

4 Implementation, Evaluation and Results of Housing Price Prediction Models

4.1 Introduction

This section discusses the implementation of ten machine learning algorithms that were used, the first five are the regression models while the last five are the classification models. For each of the models, the rationale for selecting each algorithms would be discussed in details, alongside their respective feature engineering implemented before the algorithm could be modelled, also, in this chapter, the various Evaluation metrics adopted would be discussed. in conclusion, all tools, software, hardware specification and cloud computing resources that aided the execution of this project would be discussed.

4.2 Technology & Working Environment

The work environment used for the implementation is a AMD FX-7500 Radeon R7 Processor, 10 Compute Cores 4C + 6G 2.10GHz, RAM: 8.00GB, 64-bit operating system running a windows 10 Home. The software and resources used for the completion of this project are; Rstudio: this was the major tool used for the implementation of this project, it is an open source resource with top statistical and analytical libraries for the execution of statistical tasks. Microsoft Excel: this software was used to explore and understand dataset in raw form. Tableau and PowerBi: these softwares were used to graphically plot the results achieved, in order to give better insight on the performance of each model comparatively. Google Colaboratory notebook: due to the large data set we preprocessed, when iterating several matrices, the computer runs out of space, hence the need for combined cloud computing machines resources, though this service is a paid service, it helped run the millions of iteration seamlessly. Google Geocoding API: This API was used by first obtaining the API key and this API was responsible for converting address to coordinates.

Evaluation

The ten implemented machine learning algorithms evaluated their performance in an unbiased manner, based on the previous works in Chapter 2, the predictive models were evaluated measured based on Root Mean Square Error(RMSE) and Mean Absolute Error(MAE) while the classification models were evaluated based on Accuracy.

Root Mean Square Error: Root Mean Square Error (RMSE) is one of the widely accepted and most used criteria for error measurement. This criteria was also used by Limsombunc et al. (n.d.) for error measurement.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(d-r)^2} \tag{1}$$

where, r is the predicted house price, m is the sum total of houses and d is actual house price.

Mean Absolute Error (MAE): The mean absolute error was used to measure error and compare error of the executed models. This evaluation criteria was also used by Shinde & Gawande (2018).

$$MAE = \frac{1}{M}\sum_{k=1}^{M}|d-q| \tag{2}$$

where, q is the predicted house price, M is the total number of houses and d is actual house price.

Accuracy: This metric is mostly used to know how well a model was classified, by the summation of the true positive and the true negative, divided by the total number of observations.

$$Accuracy = \frac{truePositve + trueNegative}{totalObservation} \tag{3}$$

4.3 Implementation, Evaluation and Results of Housing Price Prediction Models

4.3.1 Implementation, Evaluation and Results of Support Vector Regression

This model was implemented in the Rstudio and cross validation was performed on the testset iteratively 10 times, this was done to reduce over-fitting of the model, the train function was used to train the model with the "svmLinear" method and the metric of measurement used was RMSE. The production of hyper plane by this dataset onto classes also made it easier to identify outliers Chang & Lin (2011). **Advantages of SVM:** this algorithm was chosen because of its ability to model large dataset and it is good at utilizing memory. **Evaluation of SVM:** After the algorithm was modelled and evaluated, the SVM model has a RMSE of 0.2653 and MAE of 0.5672

4.3.2 Implementation, Evaluation and Results of Random forest

This algorithm improves with the number of trees, it was also implemented on Rstudio with the train function, using the 'ranger' method. The test set was cross-validated 10 times to reduce over-fitting. This algorithm was chosen because of its ability to work with noise. **Evaluation:** the performance of this model is measured by error rate, the RMSE observed is 0.0946 and MAE of 0.3452

4.3.3 Implementation, Evaluation and Results of Lasso Regression

This model was implemented on the Rstudio and it required the installation of the 'glmnet' package, the train function was used to implement this model with the "lasso" method, to reduce over-fitting, cross validation was performed on the test set ten times. **Evaluation:** The performance of this model is measured by error rate, the RMSE observed is 0.1506 and MAE is 0.6020

4.3.4 Implementation, Evaluation and Results of Ridge Regression

This model was implemented on the Rstudio and it required the installation of the glmnet package, the train function was used to implement this model with the "ridge" method, to reduce overfitting, cross validation was performed on the test set ten times. **Evaluation:** The performance of this model is measured by error rate, the RMSE observed is 0.1506 and MAE is 0.6021

4.3.5 Implementation, Evaluation and Results of Generalised Linear Model Regression

This supervised learning was implemented on Rstudio, for the execution of this model, the dataset needs to be normal. the package required for the implementation is "glmnet" method to train the model was used. this model was chosen because of its time efficiency. **Evaluation:** The performance of this model is measured by error rate, the RMSE observed is 0.1507 and MAE is 0.6022.

4.4 Implementation, Evaluation and Results of Housing Price Classification Models

4.4.1 Implementation, Evaluation and Results of Multi-nominal Logistics Regression

The target variable consists on more than one class, hence the binary logistics regression could not be used, this dataset was used because we have a multiclass target variable. The dataset was split into two, 80% for the trainset and 20% for the test set. The multinorm function in the Rstudio was used to train the model and the model was validated with the test set, with the confusion metric showing the true positives and negatives. **Evaluation:** The performance of this model is measured by how accurate the model was based on the data of the confusion matrix, the accuracy of Multi-nominal Logistics Regression is 51.40%

4.4.2 Implementation, Evaluation and Results of Support Vector Machine

This model was implements with the 'svm' package in the Rstudio, the 'C-classification' type was chosen because of the nature of the target variable. the four kernels (polynomial, linear and sigmoid) were all experimented, but the best performing was the polynomial kernel. **Evaluation:** The performance of this model is measured by how accurate the model was according to the data of the confusion matrix from the prediction of the test set, the accuracy of Support Vector Machine(Radial kernel) is 53.83%

4.4.3 Implementation, Evaluation and Results of Random Forest

The Random forest can be used for both regression and classification, the 'randomForest' package was installed on the Rstudio for the implementation of this model. This model was chosen because of its ability to deal with large data, even though it is not memory efficient. **Evaluation:** The performance of this model is measured by how accurate the model was according to the data of the confusion matrix from the prediction of the test set, the accuracy of the Random Forest model is 69.77%

4.4.4 Implementation, Evaluation and Results of C5.0

This model also belongs to the tree family and the parameters can be optimized, the 'c50' package was installed into the Rstudio for the implementation of the model, with the 'c5.0control' was used to control the parameter tuning, during the modelling we executed only one boosting iteration. **Evaluation:** The performance of this model is measured by how accurate the model was according to the data of the confusion matrix from the prediction of the test set, the accuracy of the C5.0 model is 66.56%

4.4.5 Implementation, Evaluation and Results of K-Nearest Neighbors

The K-Nearest Neighbors was implemented on the Rstudio by installing the 'knn' package to model the normalized training data, k-Nearest Neighbors algorithm cannot model a dataset if it is not normalized, hence the reason for feeding the algorithm with a normalized dataset. K-Nearest Neighbors has an optimum number of neighbors where the accuracy does not increase if the neighbours are increased, for this project, the optimum amount of neighbours was k=200. **Evaluation:** The performance of this model is measured by how accurate the model was according to the data of the confusion matrix from the prediction of the test set, the accuracy of the K-Nearest Neighbors model is 57.47%

4.5 Conclusion

The ten machine learning algorithms implemented and evaluated were discussed along side the benefits and disadvantages of each machine learning algorithm, then the algorithms were compared based on the evaluation techniques(RMSE, MAE and Accuracy). Also, in this chapter the various tools, software and working environment was discussed.

5 Discussion and Comparison of Developed ICT Solution

5.1 Discussion

This chapter aims to discuss the results by comparison, the scope of the project, assumptions made, how the project could be reproduced, challenges faced during implementation and design decision. The design decision of this project was adopted based on the rationale that the dataset used for this project were not created, rather the datasets were gotten from reliable sources. The 2-tier design approach was used for this project, it contains the presentation tier and the logic tier. The logic tier is often considered as the back-end where all the data cleaning, merging and transformation take place and the modelling of the datasets while the Presentation tier gives understandable insight based on the hidden patterns found by the logic tier.

The scope of this work is to examine the impact of crime occurrence in Ireland housing and how it could be used in the classification and prediction of housing prices. For the reproduction of this project, the following steps should be taken, Firstly the addresses gotten from the Irish Property were geocoded to coordinates and the coordinates of the garda stations were scraped from the garda station directory website. Secondly, the shortest distance between the the nearest primary school and busstop was gotten through their respective coordinates using R studio while the crime zones were created by attaching the a house to the nearest station, hence the house would be seen to have experienced the same number of crime in its environment as that of the number recorded by the garda station. Thirdly, all missing values and outliers were removed and before the modelling of any algorithm on this project, the datasets were split in a 80% to 20% ratio which are the training set and test set respectively. As a cautious attempt to reduce over-fitting of the model each test set was validated 10 times iterative using the cross validation technique making the results from each model reliable. The Fourth step is to check the price distribution and apply a log() function to the price feature used for

the regression models, while the "price range" used in the classification model should be encoded into three classes "x < 120000", "120000 > x < 250000" and "x > 250000" as Low Cost house, Medium cost house and High cost house respectively. The final step is to implement each algorithm and evaluate them with the test set.

The assumption of this project is, any crime that occurs in Ireland is always reported to the nearest garda station. Hence, any station with high crime report has experienced a high rate of crime. The challenges faced in his project is the high computational power needed to build a large matrix a and select the shortest distance on each matrix but this was solved by implementing the code on the Google collaboratory Notebook with computing resources of 16vCPU 104GB RAM. and another challenge was the the large errors in the Mean Absolute error, this was solved by encoding the prices to implement a classification model. RMSE was normalized because the dependent variable had large figures and the initial RMSE was not in the range of previous works.

5.2 Comparison of Implemented Models

The the five implemented predictive models had considerably low RMSE but the Mean Absolute errors were higher, this shows that the models favoured the RMSE metric better, going by this, it shows that the Random forest had the best performance with RMSE of 0.0946 and MAE of 0.3452 while the support vector machine was the least performing model with RMSE of 0.2653 and MAE of 0.5672. The comparison is clearly shown in Figure 11 below.

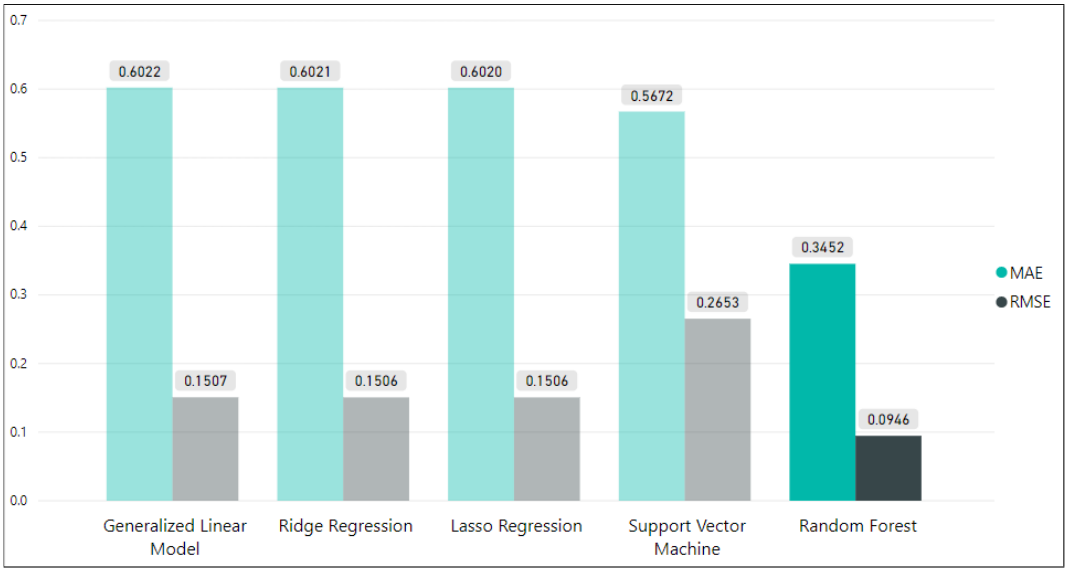

Figure 11: RMSE and MAE of the Implemented Prediction Models

Below in Figure 12 shows the accuracy of the five implemented classification models, Random forest has the highest Accuracy with of 70% while the multionomial regression showed the least accuracy of 51%.

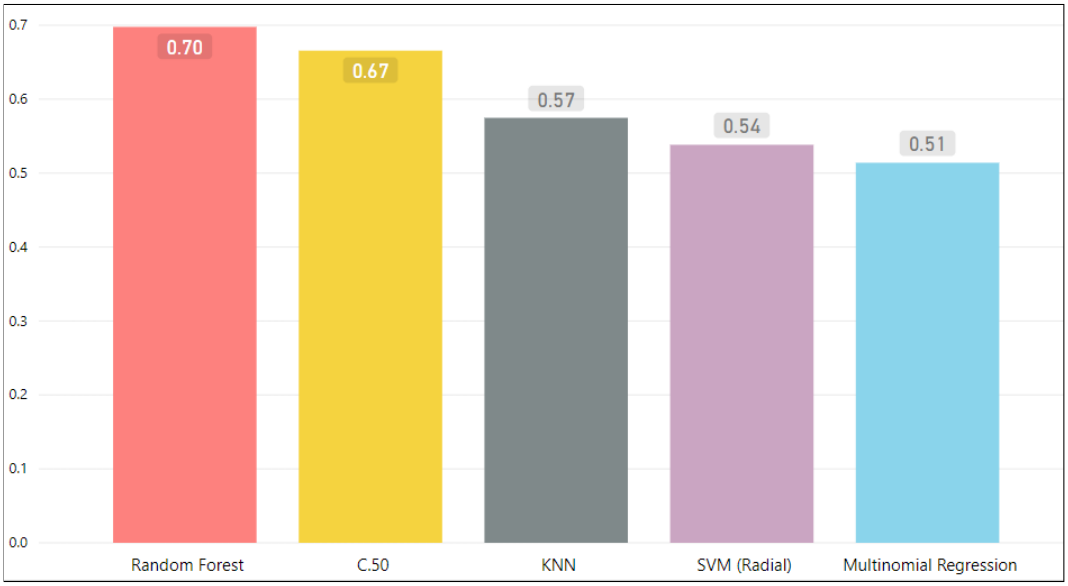

Figure 12: Accuracy of the implemented Classification Models

5.3 Comparison of Implemented Models Against Existing Models

In comparison to (Lu et al. 2017), (Yu & Wu 2016) and (Park & Bae 2015) the model had a lower RMSE result (although different datasets were used), this means the addition of Crime occurrence to the features for predicting housing prices reduced the error rate of the model also as show in Table 4 compared to the other works this project has a larger data sample size.

Table 4: Comparison of Implemented Predictive models against existing models

Author & Year	**Features**	**Sample Size**	**Result**
Sifei Lu, Zengxiang Li (2017)	Based on location, house type, size, build year, local amenities	1,300	0.1126
Yu and Jiafu Wu (2016)	House prices, numbers of bathrooms, bedrooms, living rooms, agricultural, Residential High Density, Residential Low Density, Residential Low Density Park	1,460	0.5269
B. Park and J. K. Bae (2015)	Real estate, public school ratings, and mortgage rate data	15,135	0.201
This Study	**Crime Occurrence, distance of house to Nearest Primary school and distance of house to nearest busstop**	**49,912**	**0.0946**

We compared the Accuracy of three different machine learning classification models as shown on Table 5 and it shows that this study outperformed the previous work of Yu & Wu (2016) in all the techniques.

Table 5: Comparison of Implemented Classification models against existing models

Models	(Yu and Wu, 2016.)	This Study
Multinomial Regression	0.5000	0.5140
SVM (Radial)	0.4109	0.5383
Random Forest	**0.6652**	**0.6977**

The results on Table 4 and Table 5 shows how the feature of crime occurrence in Ireland improved the prediction of housing prices.

6 Conclusion and Future Work

To answer the research question all the objectives in (Chapter 1, Subsection 1.3) were implemented, stating from the critical investigation of the Irish housing industry, price prediction models and related works within year 2002 to 2018. The five housing price prediction models (Generalised Linear Model, Ridge regression, lasso regression, support vector machine and random forest) were implemented and evaluated. The five housing price classification models (random forest, C.50, k-nearest neighbours, support vector machine and multinomial logistics regression) where also implemented and evaluated. The Implemented models were compared based on their evaluation metrics and the best performing models were compared with existing models.

This implementation has enabled us to understand the impact of of the novelty feature (Crime Occurrence) in the model and how it helps to improve accuracy and reduce error rate. According to the research question (Chapter 1, Subsection 1.2), the predictive models implemented, were Generalised Linear Model, Ridge regression, lasso regression, support vector machine and random forest with a RMSE of 0.1507, 0.1506, 0.1506, 0.2653 and 0.0946 respectively this result shows that the best performing model was random forest which had the lowest error rate and also the lowest Mean Absolute Error of 0.3452, to support the statement that the crime occurrence feature improved the performance of the model, this model was compared with three existing works of (Lu et al. 2017) who used a dataset size of 1500 and their RMSE was 0.1126, (Yu & Wu 2016) used a dataset size of 1460 with RMSE of 0.5269 and (Park & Bae 2015) used a comparatively larger dataset size of 15,135 rows with a RMSE of 0.201 but when compared to the best performing model of this project, the random forest result of RMSE 0.0946 had a better performance. To finalize the research question (Chapter 1, Subsection 1.2), the performance of the classification models random forest, C.50, k-nearest neighbours, support vector machine and multinomial logistics regression had accuracy of 70%, 67%, 57%, 54% and 51% respectively, Random forest also had the best performance with a accuracy of 70% and comparatively to the previous work of Yu & Wu (2016) who also used random forest, his accuracy was 67%, this proves that this study outperforms the existing work and in conclusion answers the research question that the crime occurrence feature helps better predict and classify housing prices.

Future Work

This project can be improved by increasing the number of physical structural features of houses and adding features like the date of initial purchase, price of initial, date of

sale, selling price, these information can be used to analyze the economic growth of each area. Also, another experiment that could be tried is to examine the impact of the price of beer on housing prices as it is a popular belief that the price of beer is correlated with the standard of living in an area. In future a larger data set can be used and a non–supervised machine learning approach can be used to create clusters of the crime in different areas, thereby creating more features.

Acknowledgement

I would specifically like to show appreciation to my supervisor Dr. Catherine Mulwa for her dedication and bespoke guide towards the success of this project. Also, many thanks to the Irish Property Price Registry, Central Statistics Office, Transport for Ireland and the Maynooth University Data Repository for making the dataset used for this project available.

References

Azadeh, A., Ziaei, B. & Moghaddam, M. (2012), 'A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations', *Expert Systems with Applications* **39**(1), 298–315.

Baumann, F. & Friehe, T. (2013), Crime and status : A contribution to strain theory.

Bin, O. (2004), 'A prediction comparison of housing sales prices by parametric versus semi-parametric regressions', *Journal of Housing Economics* **13**(1), 68–84.

Buonanno, P., Montolio, D. & Raya-Vílchez, J. (2013), 'Housing prices and crime perception', *Empirical Economics* **45**(1), 305–321.

Chang, C.-C. & Lin, C.-J. (2011), 'LIBSVM: A library for support vector machines', *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27.

Chiarazzo, V., Caggiani, L., Marinelli, M. & Ottomanelli, M. (2014), 'A Neural Network based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location', *Transportation Research Procedia* **3**, 810–817.

Cohen, M. (2008), 'The effect of crime on life satisfaction', *The Journal of Legal Studies* **37**(S2), S325–S353.

Cook, P. J. (2008), Assessing urban crime and its control: An overview, Working Paper 13781, National Bureau of Economic Research.

Espey, M., Owusu-Edusei, K. & Lin, H. (2007), 'Does close count? school proximity, school quality, and residential property values', *Journal of Agricultural and Applied Economics* **39**.

Frew, J. & Jud, G. D. (2003), 'Estimating the Value of Apartment Buildings', *Journal of Real Estate Research* **25**(1), 77–86.

Gibbons, S. (2004), 'The costs of urban property crime', *Economic Journal* **114**, 441–441.

Kahveci, M. & Sabaj, E. (2017), 'Determinant of housing rents in urban albania: An empirical hedonic price application with nsa survey data', *Eurasian Journal of Economics and Finance* **5**(2), 51–65.

Kelly, M. (2009), 'The irish credit bubble'.

Larragy, J. (2014), 'Sean ó riain 2014: The rise and fall of ireland's celtic tiger: Liberalism, boom and bust, cambridge: Cambridge university press reviewed by joe larragy for working for change: Irish community work journal may 2014', *Working For Change: The Irish Journal of Community Work* **4**(1).

Limsombunc, V., Gan, C. & Lee, M. (n.d.), 'House price prediction: Hedonic price model vs. artificial neural network', **1**(3), 193–201.

Lu, S., Li, Z., Qin, Z., Yang, X. & Goh, R. S. M. (2017), A hybrid regression technique for house prices prediction, *in* '2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)', IEEE, pp. 319–323.

Maslow, A. H. (1943), 'A theory of human motivation.', *Psychological Review* **50**(4), 370–396.

Metz, B. E. (2015), 'Effect of distance to schooling on home prices', *The Review of Regional Studies* **45**(2), 151–171.

Mukhlishin, M. F., Saputra, R. & Wibowo, A. (2017), Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor, *in* '2017 1st International Conference on Informatics and Computational Sciences (ICICoS)', pp. 171–176.

Nguyen-Hoang, P. & Yinger, J. (2011), 'The capitalization of school quality into house values: A review', *Journal of Housing Economics* **20**, 30–48.

Park, B. & Bae, J. K. (2015), 'Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data', *Expert Systems with Applications* **42**(6), 2928–2934.

Regling, K. & Watson, M. (2010), 'A preliminary report on the sources of ireland's banking crisis', p. 49.

Rosiers, F., Thériault, M., Kestens, Y. & Villeneuve, P. (2002), 'Landscaping and house values: An empirical investigation', *Journal of Real Estate Research* **23**, 139–162.

Shinde, N. & Gawande, K. (2018), Survey on predicting property price, pp. 1–7.

Varma, A., Sarma, A., Doshi, S. & Nair, R. (2018), House Price Prediction Using Machine Learning and Neural Networks, *in* '2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)', pp. 1936–1939.

Vital Statistics (2014), `https://www.cso.ie/en/releasesandpublications/ep/p-vsys/vitalstatisticsyearlysummary2014/`. [Accessed; 09-December-2019].

Yu, H. & Wu, J. (2016), 'Real estate price prediction with regression and classification', p. 5.