Assessment of Irish Mortgage Arrears at County Level using Machine Learning Techniques and Open Data

Piush Vaish

Abstract— **This paper addresses the challenge of predicting mortgage arrears at the county level for Ireland. Unlike the other works literature which focus on individual mortgages data about borrowers or exploratory analysis of lender's data, the paper presents both supervised and unsupervised data mining techniques on openly available data. The extensive experiments conducted for the paper show that it is possible to distinguish neighbouring counties from counties which are away from each other, classify provinces and predict the amount of annual arrear using derived factors from economic and census data. The benefits include an addition to the academic literature regarding arrears and the use of features and model by local authorities, lenders or the government to successfully predict the arrears and plan for the future.**

I. INTRODUCTION

In 2008, the economy of Ireland suffered a major financial crisis. The biggest drivers were the construction trade and cheap lending by the banks. The recovery since 2013 is again fuelled by growth for domestic demand, increase in house prices, construction boom and availability of mortgage. The fragile recovery is still vulnerable to a sharp decrease in the economic conditions [1]. Therefore, there is a need to learn from historical data to safeguard the economy and prevent the crisis happening again.

As a society, there are many advantages of owning a house such as reduced crime rate, stability for the family, improved educational performance, high civic volunteering and less dependence on social welfare systems. The disadvantages associated with not being able to pay the mortgage and voluntarily surrendering the house include increased homelessness, negative mental health, financial loss to the banks, increased pressure on public assistance programs, increased regulations by the central bank to cope with unexpected losses or bailout of financial institutions by the government. Thus, positive performance of the mortgage market is important for stability.

Analysis of the Irish mortgage market in 2014 indicated that only 7% were first-time-buyers. The average mortgage term was 31 years and 68% of the borrowers were under 40 years old[2]. In 2017, there was an average of 60,000 per borrower and estimated Loan-to-Value (LTV) of 90% for long-term arrears. The average balance in long-term arrears was 20% of the estimated property value. Although Dublin had the most mortgages, border and midlands were most affected by long-term arrears. For example, Cavan and Leitrim had 5% long-term arrears cases. The median interest rate was 3.7% with an average repayment amount of 848. In 2018, the average age of the borrower was 49 years and an average of 15 years was left for loans to mature [3].

In June 2018, there were 725,693 private residential mortgage accounts with a value of 98.2 billion. Of this total stock, 9.16% were in arrears. Moreover, 6.3% accounts were in arrears of more than 90 days and totalled 9.5 billion [4]. This can become a huge problem if the mortgage arrears are not controlled.

Although, the economic recovery has resulted in the pace of amount overdue to slow down. The issue is still prone to various factors such as a rise in interest rates, global economic shocks and changes in the circumstances of the borrowers to sustain payments. The recovery has also not benefited every region of the country with borrowers from South-East, Midland and Border region more prone to defaulting [1]. Arrears endanger soundness of bank's balance sheets and stability because of reduced income. It also damages the household's future creditworthiness. This can further increase volatility in future consumption, social stigma and inability to move to cheaper dwelling resulting in an economy to recover long after the recession. Unlike Gross Domestic Product (GDP), arrear is a good metric to analyse households economic hardship [5]. Hence, there is a need to provide accurate and timely insights to make sound decisions by businesses and the government.

The review of the literature shows that most of the papers focus on individual mortgages or the effect of default on the country using transactional data about borrowers. The data mining modelling techniques include regression, decision trees (DT), Gradient Boosted Classifiers (GBC) or neural networks (NN) with accuracy or precision as metrics for the classification tasks. However, there was no literature to predict the effect of mortgage arrears at the county level. This project investigates arrears by using machine learning techniques on publicly available open data.

The methodology used for the project is Knowledge discovery in databases (KDD). This paper begins with understanding the domain by researching the relevant prior knowledge and understand the user's viewpoint through literature review. It follows by discussion of steps of creating the dataset, cleaning and processing the dataset, reduction of the dataset by finding useful variables either through dimensionality reduction techniques or empirical method. Subsequently, it presents matching the goal to a data-mining method such as regression or classification, exploratory analysis and hypothesis selection which include choosing the data-mining algorithms and parameters for the model, explanation of data mining methods such as machine learning or statistical techniques to find pattern of interest from the representational data and interpreting the pattern to find meaningful interpretation. It concludes by documenting and reporting the pattern to interested parties.

II. Literature Review

Step 1 of KDD Process (Understand the domain)

A. Important Features

The careful and in-depth study of factors or variables is important to gain domain knowledge for the project and create a good quality dataset to build models. The literature suggests that "ability-to-pay" and "equity" are two most important factors to predict if a borrower will be delinquent. "Ability-to-pay" is highly correlated with either income or payment shock while negative "equity" results in the inability of the borrower re-mortgage or change the contract terms. Negative equity is defined as when the value of the property drops below the balance of the mortgage. It affects combined demand and supply of houses and the fiscal stability of the wider economy.

A study researching cross-country and within-country differences for delinquency states that factors for mortgage arrears can be group into four groups namely macroeconomic, macro-prudential regulation, institutional and housing market. It states that macro-prudential policies have become more important in most countries and need micro-sensible supervision because the tightening of lending leads to reduced housing credit growth and house price inflation. Quality of institution and the arrangement between bank and borrower affect arrears. Loan's maturity, interest rate type, tax deductibility of interest payment and legislation such as for instance that allows lenders to claim borrower's assets can also affect the number of defaults. An increase in interest rate does not have an immediate effect on delinquencies. Moreover, certain borrowers decide to default voluntarily when a household faces affordability problem[6]. Another cross-country study claims that higher mortgage debt-to-income ratio results in financial strain for the households when there is an income shock associated with either job loss or illness. Change of interest rate leads to the payment shock leading to the reduced ability of a borrower to pay. Income shock has the largest effect on the borrower's ability to pay[7].

The important factors from research regarding mortgage arrears in the Irish market ([8]; [9]; [10]; [11]; [1]) can be summarized as follows:

- Borrower specific variables such as age, marital status, number of children, occupation, monthly net income, work experience, job status. These factors are important to determine repayments with buy-to-let mortgages most likely to go into arrears than owner-occupiers.
- Mortgage specific variables such as interest rate flag, flag for more than one loan, adjustment for a loan, the age of loan, monthly payments, balance left, original amount, number of guarantors,
- Property specific variables such as the owner or buy-to-let, over or under evaluated, loan-to-value, present value, first hand or second hand, number of rooms.
- Macroeconomic information remains constant for a loan but can affect the geographic regions e.g. unemployment per region, yield curve, house price index.

Macroeconomic shock has the most impact on the borrower to default. Unemployment, income, prices of the house and loan-to-ratio (LTV) value affect the arrears at the regional level. For example, Dublin and mid-east fared better than midlands and borders across the 8 regional levels when the rate of unemployment rose 4 times between 2007 and 2010 [1]. An analysis of socio-economic factors presents that a young family head, often female, based in the urban area, recent mortgage and a long repayment term left on the loan is most likely to default in Ireland [12].

In the UK, borrower characteristics such as self-employed, first-time buyer or home-mover are important. Residential properties are more impacted by macroeconomics than behaviours of either the household or lender [13]. Study in the Spanish market uses econometric analysis to find out that "unemployment shock and fall in disposable income results in the financial vulnerability". Spain like Ireland had a crash of housing bubble in 2007 resulting in well-being of the family, homelessness and the financial system had to be rescued by the European Union [12]. Another study from Spain used the household survey to determine that younger, poorer and less well-educated households are most likely to go into arrears [14].

Monetary policy affects the arrears as a side effect. The study asserts that affordability, the logarithm of disposable income and low mortgage costs results in fewer defaults. Age also plays an important role with older households less likely to go into arrears than 35 to 44 years old. Further, large households with little education default frequently [5]. Another study states that marital quality is important to prevent arrears with wives satisfaction in a marriage has a positive effect while divorce has a negative impact [15].

Moreover, researching the effect of macroeconomic and bank-specific factors on Greek mortgage market shows that borrower and interest rates are important factors. Mortgages are normally provided to public-sector and high-skilled workers. The paper also states that delinquency decreases during the boom because of employment and a steady stream of income. This also leads to an increase in banks extending credit to low-quality borrowers [16].

Belgian market shows a similar pattern of defaults for houses facing high income or asset-to-debt ratio. Young, low-income households and often single parents are often at risk of being in debt [17]. Analysis of the survey of income and wealth from Italian households' states areas are heterogeneous in terms of economic structure, household characteristics and income distribution. Income inequality negatively affects the household's ability to make repayments and households in richer region are more likely to default they are in an unequal an area [18].

In summary, the factors for mortgage arrears are similar across Europe and can be categorized into borrower, mortgage and property specific and macroeconomic information.

B. Dataset

Data is a pre-requisite for building models. The data sets used by the works of literature can be divided as follows:

- Private data for a certain period from a lender based in Ireland for example a study used the transactional data of mortgage accounts from lender A with the employment and salary amount as borrower's information [8]. Another study used the click-stream data from a website of lender A. The website was designed to understand the customer's interaction [19] . However, in both studies the data was historical and not updated recently. Certain customers also had accounts with lenders apart from lender A. Hence, there was an incomplete picture of spending, saving and mortgage transaction habits of borrowers.
- Loan-level data collected by the Central Bank of Ireland as part of the stress-test exercise from Irish lenders such as Allied Irish Bank, Ulster Bank Ireland, Bank of Ireland (BOI) and KBC Bank Ireland (KBC). The snapshot of the data for a certain period included characteristics about the borrower, property and mortgage in addition to the information on outstanding balance and repayment terms [11],[9],[1],[3].
- Survey of financial behaviour of the household [17],[18],[14].
- Publicly available loan-level data from The Federal National Mortgage Association or Fannie Mae in the US. The single-family, fixed rate mortgage dataset contains performance and acquisition file and is used to predict the outcome of loan namely default, prepaid or paying [20],[21],[22]

In summary, there are a variety of datasets used for predicting arrears. Yet there are no publicly available datasets from the Irish lender or the Central Bank of Ireland. This results in a lack of transparency and reproducibility of the results as well as inhibit building innovative algorithms.

C. Data mining techniques

Most of the literature reported that the predictive performance of nonlinear and non-parametric algorithms is better than the traditional logit model. The imbalanced datasets were pre-processed using Synthetic Minority Over-sampling Technique (SMOTE). A study discovered that Decision Tress(DT) and AdaBoost models perform better than the other algorithms such as Logistic Regression (LR) and kNN (K-Nearest Neighbour) after addressing the imbalance in two classes using SMOTE sampling but at expense of false positive rate [8]. LR was used as a baseline by another study to benchmark the performance of models which are flexible for complex data structure and scale well depending on the size of the data. It found that Boosted Regression Trees (BRT) followed by Generalised Additive Models (GAMs) performed better than LR through empirical comparison approach. GAM's non-linear and semi-parametric property and BRT's ability to capture factor interaction resulted in good performance [11].

Another research found that Gradient Boosted Machines (GBM) is one of the top three algorithms when performed on 115 UCI binary classification datasets from 14 families of algorithms. However, none of the datasets contained mortgage data. A different study found that a mix of ensembled DT and boosting classifier are the best performing algorithms for predicting arrear but effectiveness of these models is irregular or inconsistent over time [20].

Additionally, a study states that non-linear techniques are necessary when there is a large number of variables but a linear process is suitable for small number of variables [23]. Furthermore, a study used a model on factors such as debt-to-service ratio, negative equity and unemployment rate and Root Mean Square Error (RMSE) as a metric to predict aggregated mortgage arrears in a period of stability without any major further policy interventions [24]. While another work of literature found that standard models such as LR, SVM, k-NN and NB performed poorly compared to ensembled models to predict defaults and using recall as the accuracy measure. NB was the worst because it focuses on a high false positive rate. However, the result does not consider that the data is not integrated for the lender and its subsidiary which offers mortgage under its own name. The paper suggests using the output of the default prediction models as inputs to predict arrears [19].

Another study used the Lasso approach as a benchmark to compare six approaches on AUC and RMSE criteria. The six approaches are GBM, RF model and 4 Deep leaning models with optimization of parameters on data from a European bank. It found that choice of variables and criteria produce different outputs and hence it is essential to analyse the result in detail. Tree-based models have high performance on binary classification compared to deep learning. It also uses SMOTE to over-sample the minority class by creating "synthetic" examples [25].

Furthermore, a study confirms that machine learning methods such as KNN, SVM, RF, and Factorization Machines perform better than LR when applied on dataset of 15 years on Fannie Mae dataset to predict defaults. It states the variables are important for machine learning algorithms. It discovered that loan age, LTV, credit score are important features to predict defaults [22]. A study states that linear regression and other econometric models such as generalized linear regression are normally used to find the causality for variables from theoretical knowledge when predicting arrears. The metric used is R-square. It also analysed out-of-sample accuracy by assessing training and test sets using cross-validation for models such as LR, NB, RF and kNN. However, it did not include information about borrowers such as sex, income and employment [21].

Most of the literature used AUC, precision, confusion matrix, recall and accuracy to compare the classification algorithms. One study used H-measure as an evaluation metric because the dataset was imbalanced. H-measure gives more weight to correctly classified default labels than to incorrect classified non-default labels. It still used AUC for model selection and tuning [11].

In summary, certain algorithms such as GBM and tree-based models perform well to predict defaults. Linear regression is also used to explain the variation in the features. However, the algorithms and metrics are highly dependent on the data. Any changes in the data such as behaviour of the borrower or financial sector policies can result in ambiguous result.

In conclusion, most of the papers primarily focus on individual mortgages and the effect on the country except two analysing the arrears for different regions of Ireland. The papers used transactional data from borrowers who were customers of banks. Only a few papers used loan-level data with macroeconomic variables. The important factors are the borrower, mortgage and property specific combined with macroeconomic information. Different modelling techniques such as regression, decision trees, GBM or neural networks with sampling techniques were applied. The metrics mostly used were accuracy or precision for the classification tasks.

There was search made to understand how the value of the arrears is predicted by the central bank or the lender. However, there was no literature available as predicting arrears can be an internal process used by the lenders, local authority or the central bank. Further, none of the literature used data mining to predict the effect of mortgage arrears at the county level. The data used is not open and hence further research cannot be conducted using the same dataset. This paper investigates arrears by using machine learning techniques on publicly available open data.

III. Experimental Setup

The processing, analysis and modeling were conducted using Python 3.6 and the following specifications:

- Intel(R) Core (TM) i7-6700HQ CPU @ 2.60GHz – Central Processing Unit (CPU)
- 64 GB RAM
- 4 Cores
- 8 Logical Processors
- Microsoft Windows 10 Operating System

IV. Design/ Methodology

The methodology used is Knowledge discovery in databases (KDD).

“KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” [26]. KDD is an interactive and iterative process of discovering useful knowledge from data with many decisions made by the user. It includes data mining as a step. Extracting a pattern involves fitting a model to a data while keeping the problem in mind. The process requires the models or algorithms to be able to scale to massive datasets and run efficiently. A pattern can be discovered even in randomly generated numbers. It results in discovery to be meaningless and patterns to be invalid. Hence, a valid pattern is specific to the problem and understandable with a degree of certainty. KDD has database as an important part and offers a framework to automate data analysis and hypothesis selection.

Step 2 of KDD Process (Create Target Dataset)

The dataset has been created by collecting data which is publicly available and open. These datasets are published by public sector bodies and available online as part of the Open data strategy by the Irish Government. The data offers opportunities for engagement and research to improve efficiency to deliver economic and social benefits. The data sources include data.gov, Federal Housing Finance Agency, Department of Housing, Planning and Local Government, Central Statistics Office and Property Services Regulatory Authority.

The data was put into an Analytical Base Table (ABT). ABT is a flat table containing rows and columns and contains the features for prediction. Each row contained a value of the feature per county and target feature for both classification and regression problem and no target feature for clustering [27].Before data was gathered, the appropriate design for ABT was discussed and different predictors were selected which were mentioned in the literature review as important to predict arrears. The data gathered is from either 2011 till 2017 or from 2011 and 2016. The tables for the data include:

- National average mortgage interest rates (2011-2017)
- House Prices for both new and second-hand properties (2011-2017)
- Mortgage Arrears by county (2011-2017)
- Loan Paid by an individual for both new and second-hand property (2011-2017)
- Population by gender and age group (20112016)
- Migration Population (2011 and 2016)
- Repossessions (2011 and 2016)
- Loans approved (2011 and 2016)
- Marital status such as married, divorced, single, first marriage (2011 and 2016)
- Employment status such as employed, student, retired (2011 and 2016)
- Commuter details such as children at school aged between 5 and 12 years and population aged 15 years and over at work (2011 and 2016)
- Principal economic status of the population living in the counties (2011 and 2016)
- Average Number of Persons per Private Household by County (2011 and 2016)
- Aggregate Town or Rural Area, County of Usual Residence, Nationality (2011 and 2016)

Step 3 of KDD Process (Data Cleaning and Pre-processing)

The data gathered was pre-processed to be suitable for ABT. The data quality issues such as outliers, irregular cardinality and missing values were also noted for both valid and invalid data. Invalid data is when the errors occur while creating an ABT e.g. due to calculating derived features. Invalid features are corrected to improve the quality or removed if the features are meaningless. Valid data is domain specific and require action before building models [27]. The data preparation included:

- Cleaning of the Property Register data using “MID” function in MS Excel to extract date for pivot table and filter data for each year

- Teach/rasn Cnaithe Athimhe renamed as Second-Hand Dwelling house /Apartment
- Teach/rasn Cnaithe Nua and Teach/?ras?n C?naithe Nua renamed as New Dwelling house /Apartment
- Sort the male and female population, migration, arrears and marital status data from each county in an alphabetical order
- Merged Cork City and Cork County details into one and named the column as Cork
- Merged Dublin City, Dn Laoghaire-Rathdown, Fingal and South Dublin into one and named the column as Dublin
- Merged Galway City and Galway County as one and named the column as Galway
- Interest rate was aggregated from 12 months to 1 year.
- Created a column for the four provinces of Ireland and named it as provinces.

The shape of ABT was 27 rows and 284 columns. 26 rows represent each county and one row is for the whole of the country. Hence, the dataset was small and wide and required special consideration. These considerations are discussed in the data mining (IV) section. The data was mostly numeric that is values on which arithmetic operations can be performed. County, provinces were the categorical variables as these do not allow arithmetic operation or ordered. The data quality issues were noted and acted upon. For example, while gathering data, tables which contained over 90% missing values were not used. Summary statistics such as mean, standard deviation and percentile were generated for the variables to find outliers and gain more understanding about the spread. The data for Dublin usually had values that were placed far away from central tendency than the rest of the counties. However, clamp transformation was not used to remove Dublin as an outlier because of its unique characteristics in Ireland and its interesting and predictive point of view for modelling [27]. There were no invalid outliers as the data gathered from the sites were assumed to be true. The missing values were not dropped for unnecessary loss of data or derived because ABT contained no missing values. There are a different number of counties in each province hence, there is an imbalance for classification problem. To address this issue, the metric selected were precision, recall, AUC and F1-score to measure the performance of the algorithm. For regression, each row represents a county and therefore prediction was made for each row.

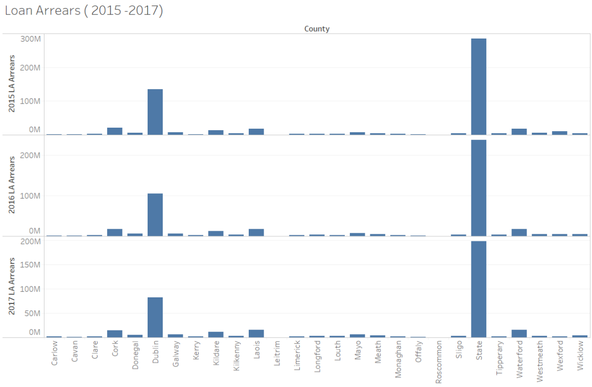


Fig. 1. Loan Arrears for County (2015 - 2017)

Step 4 of KDD Process (Data Reduction by Finding Useful Features)

The features of ABT can either be raw from the source and derived (engineered from one or more raw sources). There are numerous ways to create derived features, but the common types are aggregate e.g. average or minimum, flags e.g. indicator for absence or presence of a trait, ratio to capture relationship between two variables and mapping to chunk continuous features into categorical values. The derived feature help in reducing the number of features [27].

101 derived features were created and the type include:

- Aggregate by calculating the increase or decrease of the value of a feature such as difference of loan arrears between two years, sum of certain features such as addition of all the different nationalities living in the county for a certain year.
- Flag such as using dummy variables for different counties
- Ratio between two features such as a ratio between the number of people who commute and the total population in a county

The shape of ABT increased to 385 columns and 27 rows. The data quality issues were again reviewed and corrected for the whole dataset. The number of variables selected for the modelling was determined in different iterations, using different goals of the data mining method and included methods such as Principal Component Analysis (PCA) for dimensionality reduction, feature selection using variance threshold, feature importance using random forest, recursive feature elimination (RFE) and through greedy approach. For further explanation see data mining (IV) section.

Step 5 of KDD Process (Matching the goal to a Data-mining Method)

The aim is to predict the mortgage value of each county in a year. This goal of prediction can be achieved through a data mining method known as regression. Regression involves learning a function that maps the dependent variable to the independent variables [26]. However, the small dataset presented some challenges with having too many independent variables and a few target variables. It required a methodical process of building different parametric models using statistical test and bootstrapping with different variables. The models from literature review also suggested using simple models such as linear regression or regularised linear regression models. The suitability of the variables was validated by classifying the provinces using mortgage arrears data and predict the four provinces of Ireland. Clustering was also done on same sets of features as used for predicting the arrears to identify a set of clusters and discover hidden patterns [28]. For further explanation see IV section.

Step 6 of KDD Process (Exploratory Analysis and Hypothesis Selection)

Before building the model, an exploratory data analysis is done to understand the data characteristic such as the type and range of the values. The exploration process describes the trait of a feature using statistical measure of central tendency e.g. mode, median or mean, variation such as standard deviation or percentile and cardinality. The visualization is done to determine if features follows a certain distribution or pattern [27]. An investigation of relationship between two features was also conducted using visualisations.The two features can either be a pair of continuous or categorical variables or a continuous and categorical variable. This helped to find the features which were good for predicting the target feature and closely related features. Only one feature was included to reduce the size of ABT when there were two closely related features.

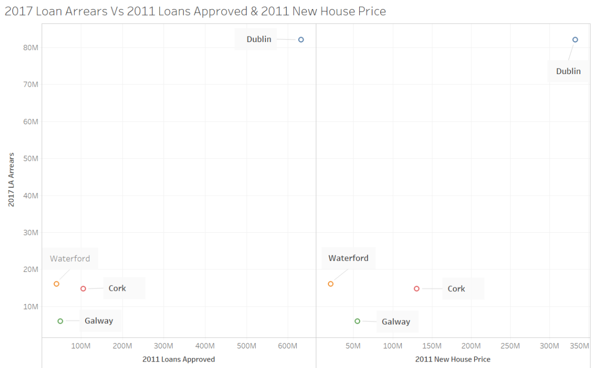


Fig. 2. Loan Vs House Price

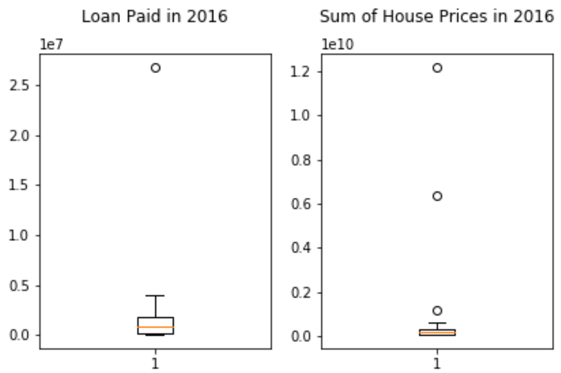


Fig. 3. Boxplot

Step 7 of KDD Process (Data mining)

KDD process involves multiple loops within an iteration or multiple iterations. All the steps are important and not only data mining for successful application in practice. The steps involve repeated iterative application of particular data-mining methods and have three primary components:

1) Model representation includes the language to find patterns while comprehending the assumptions in a method
2) Model evaluation is the quantitative value of how well the model fits.
3) Search is the optimization task to find the parameters to optimize the evaluation criteria [26].

There are certain considerations when dealing with a small data set. There is a need to keep the model simple with small set of hypotheses e.g. a decision tree with a depth of less than 3 or linear regression with 4 features. This results in avoiding overfitting the model by providing few degrees of freedom. The use of regularization also constraints the model by reducing the degree of freedom without reducing the number of variables. The assumptions should be noted and be strong e.g. lack of interaction between features to help reduce the feature space. The data can be pooled by building on top of universal model [29]. "Curse of dimensionality" states that the data needs to grow exponentially if the numbers of features are added to avoid overfitting and maintain coverage. This results in choosing the features to reduce the high dimensional space. The least discriminative features can be removed if the features are statistically independent using greedy feature selection approaches. If the features depend on one another, a single feature can represent a combination of multiple features [28]. Models with few important features are easy to interpret, reduce variance and computational cost.

The features of the model were selected using feature selection, feature extraction and empirical techniques. There were multiple iterations of model using the selected features to classify the provinces and predict the arrears. Unsupervised learning was also done using different clustering techniques to cluster regions which exhibit the same characteristics. This paper describes the experiments which achieved the best result. However, during the data mining step there were other experiments conducted with different features and modelling techniques. These were deemed unsuitable for this paper either due to the undesired outcome or statistically insignificant.

Common Approach: Splitting the data into training and test sets and normalizing the data were applied to both classification and regression modelling techniques. Train/Test Split is usually to split the data into training and test set. Training set contains the target label. The model learns from training data and makes prediction on the test data. The experiments were conducted using leave-one-out cross validation. It involved splitting the data into train/test so that each sample is used once as a test set while the remaining samples make prediction on each data point as a training set to allow. It is unbiased because the size between the whole data set and training set used in each fold is a single pattern. It is normally used on small datasets and is computationally inexpensive for certain models such as nearest- neighbour classifier, linear regression and most of the kernel methods. Standard scalar help to standardize the data by subtracting the mean and scaling to unit standard deviation. The score is calculated as $z = (x - u)/s$. It helps a dataset to look more or less like a standard normally distributed data [28].

Feature Selection: One of the experiments involved selecting features through recursive feature elimination (RFE), variance threshold (VT) and RF. RFE chooses features recursively using an external estimator and pruning least desirable features to create a smaller set. Logistic regression was used for RFE to assign weights for the top features. VT is suitable for unsupervised learning as it does not require the desired output. It picks features by removing low-level variance below the threshold. The threshold was set at 0.90. RF use tree-based strategy to rank the feature according to the improvement made in the purity of the model. The top of the tree starts with nodes which decrease impurity the most with nodes which decrease the least impurity occur at the end of the trees. Thus, a subset of important features is created by pruning trees below a particular node [27].

Experiment 1 for Classification: The top three features selected by the methods were the difference of the population who were single between 2011 and 2016, difference of departure time for students at school or college aged between 13 and 18 years between 2011 and 2016 and the difference of the persons employed for the time period between 2011 and 2016. The data was put through standard scalar before modelling. The models used for classification were the same as suggested by the literature review. The models were Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Nearest Neighbour (kNN), Decision Trees (DT), Gaussian Nave Bayes (NB), Support Vector Machine (SVM), Random Forest(RF), Gradient boosting classifier(GBC) and Multi-layer perceptron (MLP). The result of the first classification experiment is discussed in V section. Subsequently, unsupervised learning approach such as dimensionality reduction and clustering was taken to reduce the number of features.

Principal Component Analysis (PCA): PCA reduces the dimensionality by projecting the data onto a vector in order to minimize squared projection error in all directions(Witten, Frank and Hall, 2011). The following features were a good representation of the data set and therefore selected for further experimentation.

1) Sum of the differences of the loan arrears from 2011 to 2017
2) Sum of the differences in the number of mortgages from 2011 to 2017
3) Total forced repossession from 2012 to 2016.
4) Total voluntary repossession from 2012 to 2015.
5) Sum of the differences of persons employed for 2011 and 2016
6) Sum of the differences in the marital status e.g. married, single, widowed, same-sex partners, single, widowed separated for 2011 and 2016
7) Sum of the differences of the population for 2011 and 2016,
8) Sum of the differences between all kind of commuters for 2011 and 2016

The features were extracted using PCA from the above features. The implementation of PCA used was the singular value decomposition. It is suitable for dense arrays and for small dimensional data [30]. The result of PCA is discussed in V section.

Experiments for Clustering

Clustering is a way of unsupervised learning. The input data is separated on distinct groups based on similarities. Elbow method find the optimal number of clusters through explaining the percentage of variance as a function of the number of clusters so that adding more clusters do not improve the model [31] Figure 4 determine that 2 was the optimal number of clusters for k-means clustering. K-means clustering is a top-down approach and data is grouped based on the closeness to the value of the cluster.It is simple and defined the number of clusters using elbow method. Agglomerative clustering or hierarchical clustering does not require the number the clusters to be specified. It is a bottom-up approach with the points in a cluster are such that the distance between points within a cluster is minimum and distance between clusters is maximum. Affinity Propagation also does not require the number of clusters to be estimated and has no assumption on the internal structure of the data. Gaussian Mixture Modelling is probabilistic based clustering with clusters based on Gaussian distribution of the centres [30]. The result of the experiment is discussed in V section.

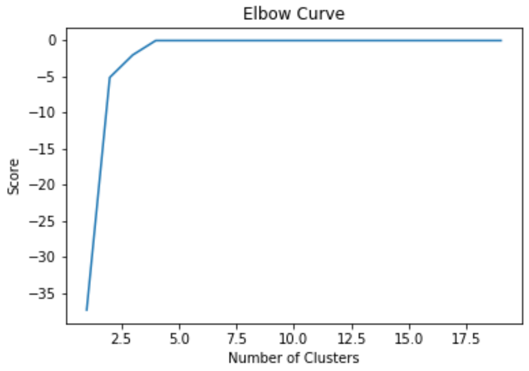


Fig. 4. Elbow Curve

Linear Regression

The mortgage arrears were predicted using regression techniques such as multivariate linear regression, Lasso regression and Ridge Regression. It is a statistical method which offers summarization and study relationship between quantitative features. The metrics used were R-squared (coefficient of determination), Adjusted R-squared and Mean Squared Error (MSE). R-squared is the “proportional improvement in prediction from a regression compared to the mean model”. The value between 0 and 1 indicates the goodness of the fit with 1 indicating a perfect prediction. The limitation of R-squared is that it can increase as predictors are added. Adjusted R-squared measures the proportion of total variance as explained by the model. It decreases if the feature added is meaningless and increase if the extra feature is worthwhile. MSE measures the average square difference between estimated values and the predicted values. It measures the quality of the model and a value close to zero is the best[29].The result of the experiment is discussed in V section.

V. Evaluation/results

Step 8 of KDD Process (Interpreting the pattern)

For the Experiment 1 for classification (IV), RF achieved the highest accuracy followed by kNN. GBC achieved 23% less accuracy than RF. However, RF is prone to overfitting because it has a fully-grown tree with low bias and high variance. It reduces error by reducing variance and hence require a large, unpruned tree based on bagging. Further, kNN loses the notion of distance to all neighbours in a high dimensional space. GBC uses regression trees and is based on high bias and low variance. The trees are shallow and reduce error by reducing the bias through optimizing an objective function. The features selected were not representative of the whole dataset and hence more experimentation was conducted.

The experiment for PCA (IV),first principal component explains almost 99.9% of the variance while the second one account for another .00015%. Figure 5 & Figure 6 show that Dublin is unlike other counties. The adjoining counties are together. Similarly, the commuter counties are very near to the major economic counties. Hence, the two principal components explain almost all the variance between the two of them.

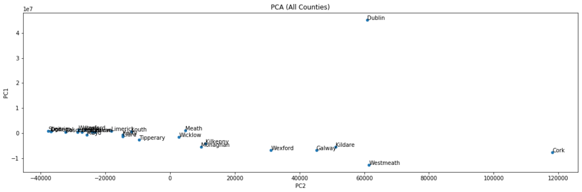


Fig. 5. PCA

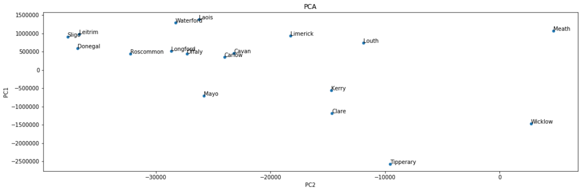


Fig. 6. PCA with few counties excluded

The four different clustering methods (IV) were implemented using the two principal components as features. However, there was no pattern discovered using clustering. It also took a lot to interpret the clusters and understand the data. The two principal components were also rejected for predicting arrears because of the lack of explainability.

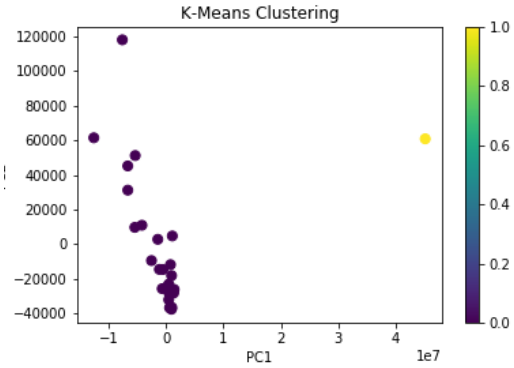


Fig. 7. K-Means Clustering

Experiment 2 for Classification: The features used for PCA were very good in determining the different counties and hence used for classifying the provinces. However, the performance of the models was not good. Elaborating the feature a level up and including finer details such as difference of loan arrears from 2011-17, difference in the population of female, employment, first time marriage, divorced, remarried, separated, single, widowed, commuter details e.g. difference in departure time of all person, children between 5-12 years, students over 19 years and going to college for 2011 and 2016, forced repossessions (2012-2016) and voluntary repossessions (2012-2015) resulted in an increase in the performance of the models. GBC predicted accuracy of 73.077% with ADA as second best at 65.385%. RF was only 42%. Hence, GBC was selected for grid search and reporting the performance metrics.

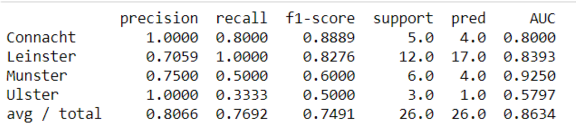
Grid search is a process of scanning the data to find the optimal parameters for a model. It is an iterative process and involves building a model on each parameter combination and storing a model for each combination. The parameters for grid search were tree specific, boosting specific and for overall use. Minimum samples split is defined as the minimum number of observations required in node for splitting. It controls overfitting and generally a small value is preferred to enable the model to learn and prevent underfitting. Minimum samples leaf is the number of samples required in the terminal node. Normally, a lower value prevents class imbalance. Maximum depth is the depth of the tree. It controls over-fitting because high depth leads to the model learning relationship specific to a feature.Learning rate controls the magnitude of the change to the estimate. Generally, a lower value ensures that the model is robust and generalizes well. Lower value also requires higher number of trees to model all the pattern. Number of estimators is the number of sequential trees. loss is the function to be minimized while random state ensures that same random numbers are generated at the same time. It is important for parameter tuning. It ensures that there is the same outcome for subsequent runs and for reproducibility [30].

Using the parameters from Table I, the accuracy increased to 76.923% . Confusion matrix is used to find the

TABLE I
PARAMETERS FOR GBC AFTER GRID SEARCH

$loss$	deviance
$learning_{r}ate$	0.1
$n_{e}stimators$	200
$min_{s}amples_{s}plit$	2
$max_{d}epth$	3
$min_{s}amples_{l}eaf$	1
$random_{s}tate$	2

correctness and accuracy of the model for two or more types of class. Precision offers information about performance regarding false positive while recall gives information about false negative. F1 score is the score that represents both the precision and recall. It is a harmonic mean between $F1Score = 2 * Precision * Recall/(Precision + Recall)$. Area under the ROC Curve (AUC) offers an aggregate metric for all possible classification threshold [28]. Figure 8 show that model has an average precision of 0.806, recall of 0.769, f1-score of 0.749 and AUC 0f 0.863. These numbers can be interpreted as on an average the model was able to catch 0.8066 classes while missed 0.7692 classes. Precision is about being precise while recall is not about catching the cases correctly but about all the cases that have a certain class. An average AUC of 0.863 means that the model has a probability of 0.863 that the model a random positive sample more highly than a random negative sample. The model was good to predict the values for the three counties namely Connacht, Leinster and Munster but not Ulster. Figure 9 shows that the departure of time to commute to work and college, marital status of a person (single or divorced), female population, the long-term loan arrears, employment status and repossessions are top 10 factors for a person living in the province. The shows the importance of work and life in Ireland. The main counties of Ireland namely Dublin, Cork, Limerick, Galway and Waterford offer benefit of scale and accumulated economies which is unique. These provide an important role for the county and the surrounding region through operating the economy and different facets of Irish social and economic life [32]. The female population is an interesting factor because it scores more important than male or total population. This can be due to the change in the attitude of women as consumers. They are also more independent and active as the Irish society changed. The Irish women have moved from the traditional role of homemaker to the dual worker in both home and labour market over the past century [33].



```
             precision  recall  f1-score  support  pred     AUC
Connacht        1.0000  0.8000    0.8889      5.0   4.0  0.8000
Leinster        0.7059  1.0000    0.8276     12.0  17.0  0.8393
Munster         0.7500  0.5000    0.6000      6.0   4.0  0.9250
Ulster          1.0000  0.3333    0.5000      3.0   1.0  0.5797
avg / total     0.8066  0.7692    0.7491     26.0  26.0  0.8634
```

Fig. 8. Performance Metrics

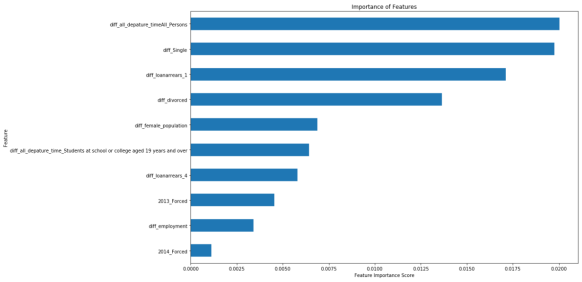


Fig. 9. Feature Importance

The features from PCA (IV) were used to find the predicted values of the arrears using multivariate linear regression and leave-one-out cross validation The results are 94.64% of r-squared value and MSE of 81,678,771,896,477.18. The feature space was further reduced to the four features:

1) Sum of the differences of the loan arrears from 2011 to 2017
2) Sum of the differences in the number of mortgages from 2011 to 2017
3) Total forced repossession from 2012 to 2016.
4) Total voluntary repossession from 2012 to 2015.

The r-squared increased to 99.3%. Lasso regression and Ridge regression offered worse performance than the multivariate linear regression even after searching for the best parameter. Hence, linear regression was tested for statistical significance. R-squared value of 99.3% indicates a strong relationship between the features and mortgage arrears. Fig. 10 shows that the predicted values are close to the actual values for most of the counties and the country. However, certain counties the predicted values are not so close. This could be that the county resulted in fewer arrears than previous years. Statistical tests were done on four individual features and as a combination. The performance of the model is more when all the features are taken into consideration.

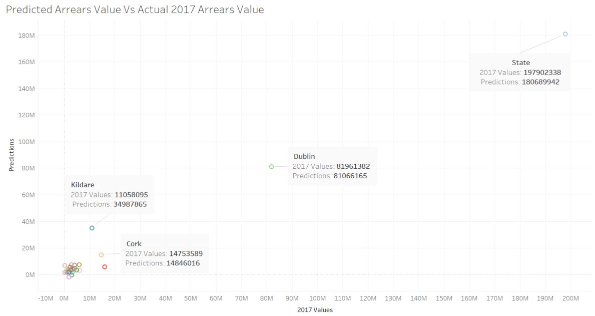
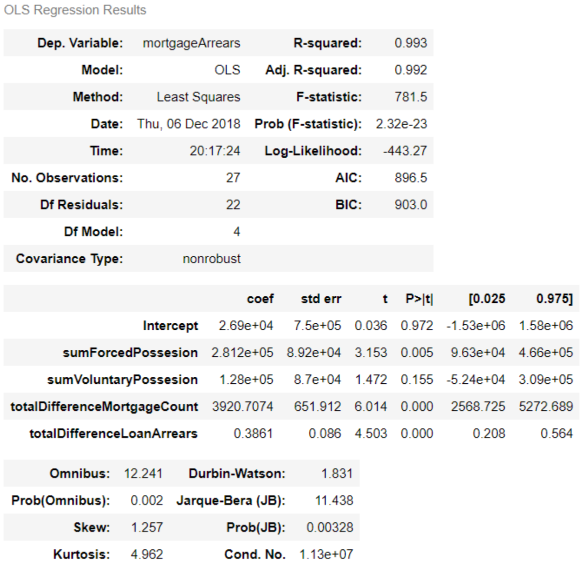


Fig. 10. Predicted Values Vs Actual Values (2017)

Figure 11 presents the linear model for the arrears as follows Arrears = 2.69e+04 + 2.812e+05 * Total forced repossessions + 1.28e+05 * Total voluntary repossession + 3920.70 * Sum of the differences of the number of mortgages + 0.3861 * Sum of the differences of the loan arrears.

The co-coefficients indicate the mean change in the arrears given one unit change in the predictor variables. Standard error is the quantitative measure of uncertainty e.g. 0.086 is



OLS Regression Results

Dep. Variable:	mortgageArrears	**R-squared:**	0.993
Model:	OLS	**Adj. R-squared:**	0.992
Method:	Least Squares	**F-statistic:**	781.5
Date:	Thu, 06 Dec 2018	**Prob (F-statistic):**	2.32e-23
Time:	20:17:24	**Log-Likelihood:**	-443.27
No. Observations:	27	**AIC:**	896.5
Df Residuals:	22	**BIC:**	903.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.69e+04	7.5e+05	0.036	0.972	-1.53e+06	1.58e+06
sumForcedPossesion	2.812e+05	8.92e+04	3.153	0.005	9.63e+04	4.66e+05
sumVoluntaryPossesion	1.28e+05	8.7e+04	1.472	0.155	-5.24e+04	3.09e+05
totalDifferenceMortgageCount	3920.7074	651.912	6.014	0.000	2568.725	5272.689
totalDifferenceLoanArrears	0.3861	0.086	4.503	0.000	0.208	0.564

Omnibus:	12.241	**Durbin-Watson:**	1.831
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	11.438
Skew:	1.257	**Prob(JB):**	0.00328
Kurtosis:	4.962	**Cond. No.**	1.13e+07

Fig. 11. OLS Regression Model

the possible difference between the true arrears and the sum of the differences of the loan arrears. The coefficient values of this feature are also the high hence it is good indicator for predicting the mortgage arrears. This relationship is like as discussed in different works of literature. F-statistics is 781.50 and Prob(F-statistics) is very small at 2.320e - 23. This low value indicates that there is a small chance that all of the regression parameters are zero and that the regression equation does have independent variables that are not purely random with respect to the dependent variable. The regression model explains 99.3% of the variation in arrears due to features. Hence, large variance is accounted for by the regression model. Consequently, the features are good for modelling[34].

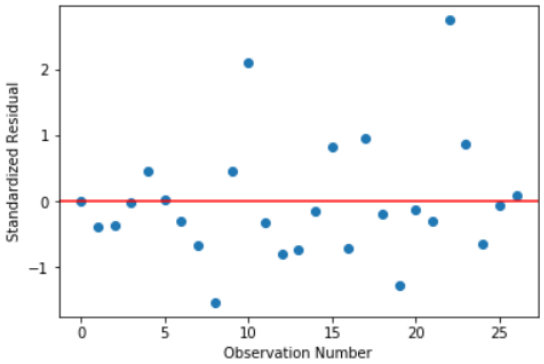


Fig. 12. Residual Plot

Figure 12 falls in a symmetrical pattern and have a constant spread throughout the range. Hence, the linear regression model is appropriate for the data. Figure 13 is a scatterplot of two sets of quantiles. The plot shows that both sets of quantiles did not came from same normal distribution and there are some extreme values in the data [35]. These values are of Dublin and State which are different than rest of the counties in Ireland.

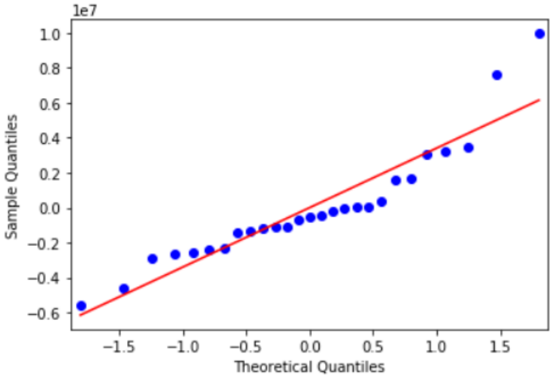


Fig. 13. Q-Q Plot

VI. Conclusion

Step 9 of KDD Process (Documenting and Reporting)

The project involved producing models for both unsupervised and supervised learning. The two principal components from PCA explains almost all the variance using total forced repossessions, total voluntary repossession, sum of the differences of the number of mortgages and sum of the differences of the loan arrears as features. The analysis helps to distinguish neighbouring counties from counties which are away from each other. Dublin shows a unique position among the various counties of Ireland. A multiclass classification is conducted for classifying the four provinces based on the open data collected for arrears. The final model with optimized parameters is able to identify the provinces with 0.76 AUC. The same features are again used to predict the value of mortgage arrears in each county and the entire state. The regression model can explain 99.3% of the variation in arrears due to these features:

1) Sum of the differences of the loan arrears from 2011 to 2017
2) Sum of the differences in the number of mortgages from 2011 to 2017
3) Total forced repossession from 2012 to 2016.
4) Total voluntary repossession from 2012 to 2015.

Hence, the features can be used for modelling and the model can be used by local authorities or the government to successfully predict the arrears.

A. Contributions to the Body of Knowledge

The paper helps to provide an understanding of the mortgage arrears with a focus on Ireland. It demonstrates that using derived factors from economic and census data can help to determine the arrears and improve the predictive capability of the model. It provides the features which are important to categorize counties in different provinces. However, further research must be conducted to ascertain the other factors which are contributing to these features. The paper shows that it is possible to build models using public and openly available data. Despite the lack of academic literature showcasing the prediction of the arrears used by the lenders or Central Bank of Ireland, the models presented can certainly be a viable alternative to some of the more complex models. It also adds to the academic research in the area with a novel and unique way not discussed in the literature. The authorities can do better financial planning for the issue going forward by using a simple baseline model and then incorporate additional data to gain more understanding.

B. Limitations

A substantial issue for the paper is the amount of data available due to a small number of counties in Ireland. Furthermore, the data collected is from a period when the economy has recovered after the recession. The models must be trained again with substantial data and few adjustments for other economic trends such as decreased consumer spending and increased job losses during recession in order to make more accurate and effective plans. Another scenario when the model needs to be updated is when the policies of the authorities are changed due to the predictions made by the model leading to a change in borrowers behaviour and activity changes. This is known as concept drift which states that the target is likely to change as the model progresses and matures.

C. Future work

Several areas can be investigated further to gain a better understanding of the arrears. The data of Dublin and State are outliers compared to the other counties. Hence, the predictions can be calculated again to measure the effect on the model after removing these data points. The model can be run on the smaller area of regions such as cities and towns in Ireland to determine the usefulness of the data. The data about borrowers from counties can be gathered and added to help improve the predictions. The addition of transactions habits of the borrower with the macroeconomic data can further provide better models for a wider economic change.

References

[1] Vasilis Tsiropoulos. *Financial Stability Notes A Vulnerability Analysis for Mortgaged Irish Households A Vulnerability Analysis for Mortgaged Irish Households*. Tech. rep. 2018.

[2] David Duffy and Niall O'Hanlon. "Negative equity in Ireland: estimates using loan-level data". In: *Journal of European Real Estate Research* 7.3 (Oct. 2014). Ed. by Kenneth Gibb and Alex Marsh, pp. 327–344. ISSN: 1753-9269. DOI: 10.1108/JERER-01-2014-0009.

[3] Terry O'malley. *Financial Stability Notes Long-Term Mortgage Arrears in Ireland Long-Term Mortgage Arrears in Ireland*. Tech. rep. 2018.

[4] Central Bank of Ireland. *Statistical Release*. Tech. rep. 2018.

[5] Petra Gerlach-Kristen and Seán Lyons. *Mortgage arrears in Europe: The impact of monetary and macroprudential policies; Mortgage arrears in Europe: The impact of monetary and macroprudential policies*. Tech. rep. 2015.

[6] Irina Stanga, Razvan Vlahu, and Jakob de Haan. "Mortgage Arrears, Regulation and Institutions: Cross-Country Evidence". In: *SSRN Electronic Journal* (Dec. 2017). DOI: 10.2139/ssrn.3094242. URL: https://www.ssrn.com/abstract=3094242.

[7] John Whitley, Richard Windram, and Prudence Cox. "An Empirical Model of Household Arrears". In: *SSRN Electronic Journal* (Mar. 2004). ISSN: 1556-5068. DOI: 10.2139/ssrn.598886.

[8] Jamie Roche. *Predicting Mortgage Arrears: An Investigation Into the Predictive Capability of Customer Spending Patterns Recommended Citation*. Tech. rep. 2014.

[9] Reamonn Lydon and Yvonne McCarthy. *What Lies Beneath? Understanding Recent Trends in Irish Mortgage Arrears*. Vol. 44. 1, Spring. [Economic and Social Studies], Sept. 2013, pp. 117–150.

[10] Yvonne McCarthy. "Disentangling the Mortgage Arrears Crisis: The Role of the Labour Market, Income Volatility and Housing Equity". In: *SSRN Electronic Journal* (Jan. 2014). ISSN: 1556-5068. DOI: 10.2139/ssrn.2536924.

[11] Trevor Fitzpatrick and Christophe Mues. "An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market". In: *European Journal of Operational Research* 249.2 (Mar. 2016), pp. 427–439. ISSN: 0377-2217. DOI: 10.1016/J.EJOR.2015.09.014.

[12] M. Teresa Sánchez-Martínez, Jose Sanchez-Campillo, and Dolores Moreno-Herrero. "Mortgage debt and household vulnerability". In: *International Journal of Housing Markets and Analysis* 9.3 (Aug. 2016), pp. 400–420. ISSN: 1753-8270. DOI: 10.1108/IJHMA-07-2015-0038.

[13] Lu Zhang, Arzu Uluc, and Dirk Bezemer. *Staff Working Paper No. 651 Did pre-crisis mortgage lending limit post-crisis corporate lending? Evidence from UK bank balance sheets*. Tech. rep. 2017.

[14] Carlos Aller and Charles Grant. "The effect of the financial crisis on default by Spanish households". In: *Journal of Financial Stability* 36 (2018), pp. 39–52. DOI: 10.1016/j.jfs.2018.02.006.

[15] Robert C Stewart, Jeffrey P Dew, and Yoon G Lee. "The Association between Employment- and Housing-Related Financial Stressors and Marital Outcomes during the 2007fffdfffdfffd2009 Recession". In: *Journal of Financial Therapy* 8.1 (July 2017). ISSN: 1944-9771. DOI: 10.4148/1944-9771.1125.

[16] Dimitrios P. Louzis, Angelos T. Vouldis, and Vasilios L. Metaxas. "Macroeconomic and bank-specific

determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios". In: *Journal of Banking & Finance* 36.4 (Apr. 2012), pp. 1012–1027. ISSN: 0378-4266. DOI: 10.1016/J.JBANKFIN.2011.10.012.

[17] Du Caju. *A Service of zbw Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics*. Tech. rep. 2017.

[18] David Loschiavo. *Temi di Discussione Household debt and income inequality: evidence from Italian survey data*. Tech. rep. 2016.

[19] Gavin O'brien. *Clicking into Mortgage Arrears: A Study into Arrears Prediction with Clickstream Data*. Tech. rep. 2018.

[20] Jesse Sealand. "Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models Application of Gradient Boosting Algorithms for Prediction of Relapse in Childhood Acute Lymphoblastic Leukemia View project". In: (2018). DOI: 10.13140/RG.2.2.30004.76169.

[21] Bolarinwa Akindaini. *MACHINE LEARNING APPLICATIONS IN MORTGAGE DEFAULT PREDICTION*. Tech. rep. 2017.

[22] Ali Bagherpour. *Predicting Mortgage Loan Default with Machine Learning Methods*. 2017.

[23] Flavio Barboza, Herbert Kimura, and Edward Altman. "Machine learning models and bankruptcy prediction". In: *Expert Systems with Applications* 83 (Oct. 2017), pp. 405–417. ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2017.04.006.

[24] Janine Aron and John Muellbauer. "fffdfffdfffdModelling and forecasting mortgage delinquency and foreclosure in the UK.fffdfffdfffd". In: *Journal of Urban Economics* 94 (July 2016), pp. 32–53. ISSN: 0094-1190. DOI: 10.1016/J.JUE.2016.03.005.

[25] Peter Addo, Dominique Guegan, Bertrand Hassani, et al. "Credit Risk Analysis Using Machine and Deep Learning Models". In: *Risks* 6.2 (Apr. 2018), p. 38. ISSN: 2227-9091. DOI: 10.3390/risks6020038.

[26] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases". In: *AI Magazine* 17.3 (Mar. 1996), pp. 37–37. ISSN: 2371-9621. DOI: 10.1609/AIMAG.V17I3.1230.

[27] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. 2015, p. 595. ISBN: 0262029448.

[28] Gareth James, Daniela Witten, Trevor Hastie, et al. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7.

[29] M Jordan, J Kleinberg, and B Schölkopf. *Pattern Recognition and Machine Learning*. Tech. rep. 2005.

[30] Scikit-learn. *scikit-learn: machine learning in Python fffdfffdfffd scikit-learn 0.20.1 documentation*. 2018. (Visited on 11/27/2018).

[31] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining*. 2011. (Visited on 11/24/2018).

[32] By Goodbody Economic Consultants. *The Role of Dublin in Europe A Report prepared for the Spatial Planning Unit, Department of the Environment and Local Government*. Tech. rep. 2000.

[33] Ashling Sheehan, Elaine Berkery, and Maria Lichrou. "Changing role of women in the Irish society: an overview of the female consumer". In: *The Irish Journal of Management* 36.3 (Dec. 2017), pp. 162–171. ISSN: 1649-248X. DOI: 10.1515/ijm-2017-0017.

[34] The Pennsylvania State University. *1.2 - What is the "Best Fitting Line"? — STAT 501*. 2018. (Visited on 11/30/2018).

[35] Clay Ford. *Understanding Q-Q Plots — University of Virginia Library Research Data Services + Sciences*. 2015. (Visited on 11/30/2018).

APPENDIX

TABLE II
TEAM MEMBER'S CONTRIBUTION

Name	Contribution
Piush Vaish	Data Gathering, Cleaning, Visualization and Data Mining
Ravikanth Dulam	Data Gathering, Cleaning, Visualization and Data Mining