Rapid Assessment of Real Estate Loan Disapproval via Predictive Modeling: A Case for the Philippines

Adrian Nicholas A. Corpuz
Joseph Ryan G. Lansangan
School of Statistics
University of the Philippines Diliman

The Philippines is currently experiencing a housing backlog and is expected to reach 6.5 million by the year 2030 if nothing is done about this. It is in this context where the government and the private sector have partnered themselves to address the backlog. Financing institutions such as private banks and the Home Development Mutual Fund (i.e. Pag-IBIG) offer different home loans for Filipinos to be able to afford these houses. Using a local real estate development's dataset, the study explores the application of predictive models in quickly determining whether a client will likely be able to get a home loan approved or not once he or she submits the preliminary documents for a home loan. Results show that in terms of accuracy, decision trees and random forest are superior in predicting home loan disapproval than binary logistic regression. The best predictive model is the random forest model, and results show that the main determinants of getting a home loan approved are loan equity term, total contract price of the house, equity payment status, and the income of the client.

Keywords: *real estate, home loan, binomial logistic regression, decision tree, CART, CTree, CHAID, random forest*

1. Introduction

Housing is referred to as a catalyst for the development of the economy, an instrument for business development and economic activity that does not only help the real estate industry but many other sectors as well such as employment, education, healthcare, and tourism. Housing is one of the primary indicators of inclusive economic growth as it addresses one of the basic human needs. Thus, poverty alleviation will be significantly lessened if the different real estate developers in the Philippines will make their residential housing products more accessible to the Filipino people. One of the main reasons of not being able to provide houses to Filipinos is their lack of capacity to pay for monthly amortizations and equities (Padojinog, et al., 2012).

While the majority of housing developers in the country practice selling house and lot packages prior to construction, there is still a risk of house forfeiture due to inability to get a housing loan approved. Predictive modelling will further help developers minimize the possibility of forfeiture due to home loan disapproval which in turn would lead to better revenues for the developers. It will also help Filipino homebuyers be able to afford homes in the long term by getting their home loans approved from Philippine financing institutions.

Interestingly enough, for the Philippines that needs to address its housing backlog of 6.5 million units by the year 2030, the country's real estate developers still fail to provide homes for Filipinos (Padojinog, et al., 2012). Data collected from Property Company of Friends Inc. (Pro-Friends), one of the country's leading developers of affordable housing, show that there are more clients who apply for bank loans and Pag-IBIG loans to pay for their house rather than in-house financing where interest rates are usually higher.

Developers including Pro-Friends have depended heavily on the Credit Management Association of the Philippines (CMAP) to regulate housing loan applications. CMAP does not explore other variables such as sex/gender, type of housing purchased, or location of house. The CMAP model is also not publicly available, hence, developing a predictive model using available data may help predict loan disapproval among clients.

Being denied a loan from a bank or a Home Development Mutual Fund (HDMF) would lead to a loan with higher interest rate (i.e. in-house financing). The development of a predictive model hopes to address the concern on higher interest rates by exploring the characteristics of home loan applicants and what gets them disapproved in their loan when applied at the period after the buyer submits his or her preliminary housing loan documents. The predictive model would help developers identify home loan applicants who are likely to be denied a home loan earlier and advise them to shift to another house model.

Rapid assessment of loan applications by residential developers is needed even before applying for a home loan (from banks and Pag-IBIG) in order to avoid house forfeiture because of the inability to find appropriate financing for their house. Affordable housing developers which provide diverse house financing options in partnership with banks and Pag-IBIG (the Philippines' version of the Home Development Mutual Fund) typically face forfeitures from some of customers. This is usually because of the low reservation fee, insufficient monthly income of clients, insufficient documents, and not getting their home loan approved after the home equity period (Alonzo, 1994). These developers have tried experimenting with different financing schemes and yet a number of forfeitures still exists. High forfeiture projects are generally seen in residential developments wherein units are already being built even without the guarantee of a buyer occupying the said units (i.e. practice of pre-selling). Predictive models relative to housing loans will help the developer provide homes to those in need without compromising revenue as well as the homebuyers getting the financial help from loans that will allow them to purchase their own house.

In this study, predictive models are developed for determining the probability of a home loan being disapproved by a financing institution using client demographic data and housing development information. Through the different models, client characteristics and/or development information most important in determining whether a buyer's loan will likely be disapproved (or not) are identified. A final prediction model useful for rapid assessment is then recommended.

In the following section, a brief discussion on the situation of the Philippine housing sector is presented. In Section 3, studies on loan applications are given. The data and methods used in the study are then presented in Section 4, and results and discussions are provided in Section 5. The final section provides concluding remarks.

2. Situation of the Philippine Housing sector

2.1. General demographics

The Philippines expects continued growth in both population and economic outlook and as a result will need more housing units from the various housing developers of the country. Population size was estimated to be 100,981,437 as of August 2015 (Philippine Statistics Authority, 2017), and the average population growth rate of the country was 1.7% annually during the period of 2010 to 2015. In other words, over this period, roughly 17 people were added for every 1,000 persons in the population. Using medium assumptions from the Philippine Statistics Authority (PSA), the Philippines will reach a population of 125,337,500 by the year 2030. The number of households in the country is expected to reach 31.5 million from the 22.4 million households in 2015, which translates to four persons for every household (PSA, 2018). Projected population and number of households by 2030 is at 125.3 million and 31.5 million, respectively (PSA, 2019).

2.2. Housing supply and housing segments

The country's major and leading organization of housing and real estate, Subdivision and Housing Developers Association (SHDA) has partnered with the Center for Research and Communication at the University of Asia and the Pacific (UAandP) to undertake a study entitled *The Housing Industry Road Map of the Philippines: 2012-2030*. This study gives an overview of the supply needed for every housing segment and the required income of Filipino buyers to be able to purchase a house. One of the highlights in the report is that while different loan terms and interest rates are offered for every segment, majority are in the socialized housing segment. Consequentially, this would mean that their housing loans should have lower interest rates and longer loan terms because the target market of socialized housing are those of lower income (Alonzo, 1994).

Those that are in need of longer loan terms and lower interest rates are in the first three sectors of the definitions used by both SHDA and the Housing and Urban Development Coordinating Council (HUDCC) (Generalao, 2017). These three sectors are socialized, economic, and low-cost housing. The sectors and their respective price ranges as defined by SHDA and HUDCC are seen in the following table.

Table 1. Housing Segment Definitions

Segment	Price Range (in PHP)
Socialized Housing	Below 450,000
Economic Housing	450,001 to 1,700,000
Low Cost Housing	1,700,001 to 3,000,000
Medium Cost Housing	3,000,001 to 4,000,000
Open Market/High End Housing	4,000,001 above

Source: Subdivision and Housing Developers Association and Housing and Urban Development Coordinating Council (2017), http://hlurb.gov.ph/wp-content/uploads/IRR/IRR_RA_9904.pdf, *retrieved June 2019*

2.3. The housing backlog

Using the data gathered and analyzed in *The Housing Industry Road Map of the Philippines: 2012-2030* by SHDA and UAandP, the total housing backlog in the country as of 2011 is roughly 3.9 million households (Padojinog, et al., 2012). Approximately 80% of the housing backlog are in the socialized and economic housing segment. Naturally, one of the major reasons why there is a need to implement RA No. 10884, the Balanced Housing Development Program, is because there is no housing backlog in the mid-income and high-end segments. Having no backlog in the latter two segments is an indicator that real estate developers have been focused on providing houses to the middle and high class of the country in the past (Padojinog, et al., 2012).

The study led by W. Padojinog in 2012 did not account for RA No. 10884. Their team, in collaboration with SHDA, computed that roughly house production will average at 200,000 units every year from 2012 to 2030 but the housing need will average at 345,941 per year. Thus, there will be a backlog of 145,941 houses annually and 6.5 million additional households will be needed by the year 2030. To further illustrate their computation, Padojinog et al. (2012) has provided the following table which shows the current housing backlog, new housing needs annually from 2012 to 2030, housing production capacity, and the housing backlog by 2030.

Table 2. Estimated housing backlog by 2030

Particulars	Units Per Year	Number of Years	Total Units
Current Housing Backlog	-	-	3,919,566
New Housing Need 2012 – 2030	345,941	18	6,226,540
Housing Production Capacity	200,000	18	(3,600,000)
Backlog by 2030	-	-	6,546,106

Housing backlog in the Philippines can still be classified as grossly underestimated since the number of poor people who live in informal settlements is not accounted for in the housing backlog (Ballesteros, 2010). Houses that do not pass the safety standards of housing and houses that are constructed on land wherein occupants do not have a legal claim should be considered in the housing backlog.

2.4. Republic Acts No. 10884, 10963 and 11201

Three government laws relative to this study are discussed. Republic Act (RA) No. 10884 is important in the context of the real estate industry mainly because it is the policy that requires all residential developers to help address the housing backlog of the country and why there is a need to provide home loans for Filipinos to afford housing. RA No. 10963 (TRAIN Law) is noteworthy because it has essentially increased prices across all housing sectors (e.g. socialized housing and mid to high-end housing) which influences the affordability of Filipinos for houses. Finally, RA No. 11201 is an act that consolidates all housing agencies in order to efficiently address housing concerns including house financing and shelter provision, it is relatively new compared to the other Republic Acts but its influence (on housing production and financing) is expected in the future because it is the designated housing arm of the government.

To address the problem of having the poor not being able to afford homes in the Philippines, the government has introduced the Balanced Housing Development Program Amendments, RA No. 10884, in 2018 to revise the original Urban Development and Housing Act (UDHA) of 1992 otherwise known as RA No. 7279. RA No. 10884, which addresses the housing backlog supply, helps majority of future homebuyers in purchasing a home in the country, mostly in the socialized and economic housing segment.

RA No. 10884 requires Philippine Real Estate Developers to develop "an area for a socialized housing project equivalent to at least fifteen percent (15%) of the total subdivision project area or the total subdivision project cost and at least five percent (5%) of condominium project area or condominium project cost, at the option of the developer" (Housing and Land Use Regulatory Board, 2018).

RA No. 10963: Tax Reform and Acceleration and Inclusion or the TRAIN Law is an initiative that the Philippine administration believes to be instrumental in reducing poverty and reaching economic development goals (Congress of the Philippines, Republic Act No. 10963, 2017). The TRAIN Law has impacts on various industries, including the business of real estate developers. Essentially, there are three different factors in the TRAIN Law that affect the real estate industry: (1) cost of building materials, (2) changes in estate tax, and (3) the exemption in Value Added Tax (VAT).

TRAIN Law increases the tax on materials needed in construction such as oil, petroleum, and coal, and as a result, has driven up the prices for housing. Changes in estate tax under the TRAIN Law subject beneficiaries of real estate inheritance to a 6% flat rate of estate tax; as reference, before TRAIN Law was passed, all properties worth P200,000 and above will be subjected to a 5% to 20% tax. It is also noteworthy that family homes certified by the LGU that are worth P10,000,000 and below will be exempted from the estate tax. The exemption in Value Added Tax greatly affects those that are buying a house in the socialized housing segment with the aim of lightening the load of real estate tax on the poor and leveling the tax on the mid-market and high-end market housing. Under the TRAIN Law, housing units that are priced P2,500,000 and below will have VAT exemptions and thus easing the load of tax payments for the socialized market (Congress of the Philippines, Republic Act No. 10963, 2017).

RA No. 11201: Creation of Department of Human Settlements and Urban Development (DHSUD), on the other hand, creates the DHSUD by merging the Housing and Urban Development Coordinating Council with the HLURB. The DHSUD is now "the primary national government entity responsible for the management of housing, human settlement and urban development" (Congress of the Philippines, Republic Act No. 11201, 2019). Under the law, the department shall have two divisions: the human settlements division, and the urban development division. In RA No. 11201, "human settlements comprise of the physical features and components of a shelter and infrastructure; and services wherein physical elements that provide support such as health, culture, education, recreation, welfare, and nutrition." On the other hand, urban development refers to the process of occupation and land use for activities covering residential, industrial, commercial, and all in between that is necessary to carry out urban living functions (Congress of the Philippines, Republic Act No. 11201, 2019).

RA No. 11201 thus abolished the Housing and Urban Development Coordinating Council and reconstituted the HLURB as the Human Settlements Adjudication Commission (HSAC). The DHSUD will have powers and functions based on four key areas. These are (1) policy development, coordination, monitoring and evaluation; (2) environmental land use and urban planning and development; (3) housing and real estate regulation; and (4) homeowners association and community development. The department is tasked with the establishment of a one-stop processing center to centralize the processing of housing-related permits, clearances, and licenses. DHSUD is attached with four government-owned corporations significant to Philippine housing – National Housing Authority (NHA), National Home Mortgage Finance Corp. (NHMFC), Pag-IBIG (or HDMF), and Social Housing Finance Corp. (SHFC).

2.5. Interest rates and bank financing in the Philippines

Interest rates of housing loans in the Philippines play an important role in providing home loan approvals to potential clients (Patnaik et al., 2017). Historically speaking, low interest rates is one of the major contributors in building a strong housing market (Tsataronis and Zhu, 2004). However, it is worthy to note that the Philippines is presently experiencing higher inflation rates. Generally, as interest rates are increased, less people borrow money and essentially, less people apply for loans. This means that when interest rates are high, supply for money is less and therefore inflation decreases (Fernando et al., 2001). Interest rates are increased by the central bank to control high inflation. Historical data of inflation shows that it has doubled in the last 5 years, starting at 2.6 in 2013 and then reaching 5.2% in 2018 (PSA, 2019).

Table 3. Home Loan Rates of Philippine Banks (2019)

Bank	1 Year	2-4 Years	5 Years	10 Years	15 Years
Bank of the Philippine Islands	7%	7.25% to 8%	8.50%	9.50%	10.50%
Banco de Oro	6.50%	-	-	-	-
Chinabank	6%	6.50%	7%	9%	-
United Coconut Planters Bank	7%	8%	9%	-	-
PSBank	7%	7.50%	8.50%	9.50%	-
EastWest	7%	7.50%	8%	8.75%	10%
Metrobank	-	7.25% to 8%	8.50%	-	-
Unionbank	8%	8.25%	8.75%	10%	-
Security Bank	6.50%	7.50%	8.50%	-	-
AVERAGE Housing Loan Rate	7%	8%	8%	9%	10%

The bank loan approach is one of the most common ways of financing the purchase of a house in the affordable to high-end market. This approach requires a collateral from the homebuyer wherein the property being bought is the collateral itself. While requirements for bank loans vary from bank to bank, banking institutions typically ask for employment documents, purpose for the loan application, and the type of property that is being bought. For example, Banco de Oro, one of the banks in the Philippines, offers home loan to applicants that (1) are at least 21 years old but not older than 70 years old, (2) has stable source of income – requirements vary if locally employed, self-employed or employed abroad (i.e. Overseas Filipino Worker), and (3) can provide documents such as government ID, marriage contract if applicable, and income documents (Source: https://www.bdo.com.ph/mobile/personal/loans/home-loan, retrieved June 2019).

Banks usually provide only up to 15 years of financing while Pag-IBIG can provide house financing of up until 30 years. For reference, interest rates in Philippine banks which range from 6 to 10.5% depending on the loan terms, can be found in the following table (Rivas, 2019).

2.6. Pag-IBIG and in-house financing

Pag-IBIG Financing is the typical home loan financing of houses that are in the socialized and economic market or housing that are priced less than or equal to PHP 1.7 million (Padojinog, et al., 2012). Pag-IBIG is a Philippine government-owned institution under the DHSUD. The institution offers short to long-term loans for housing. Generally providing the most accessible home loan financing available, Pag-IBIG only requires members to be employed, to be not over 60 years old, and must be earning at least PHP 1,000 per month. Pag-IBIG usually offers the longest loan terms of up until 30 years payment and the lowest interest rates at only 3% per annum (Franco-Garcia and Galang, Jr., 2017). To illustrate how accessible Pag-IBIG Financing is to the Filipino masses, those who apply for a PHP 450,000 loan will only need to pay approximately PHP 1,900 for 360 months or 30 years.

For reference, the interest rate of Pag-IBIG Housing Loans typically ranges from 5.375% to 10% per annum that have a fixed pricing period of up to 30 years. The fixed pricing period is very similar to the loan duration of loans offered by banking institutions. The fixed pricing period is the duration wherein how long the interest rate is locked, meaning it will not be changed or repriced within the duration of the period. For example, a 20-year fixed pricing period in Pag-IBIG will have an interest rate of 8.8% for 20 years. After the period of 20 years, the loan will be repriced and thus a higher interest may occur. The 3% interest rate per annum with a loan duration of 30 years is only applicable to loans that are less than PHP 450,000; and Filipino minimum-wage workers or those earning less than PHP 15,000 for Metro Manila workers and PHP 12,000 for workers outside Metro Manila (Franco-Garcia and Galang, Jr., 2017).

On the other hand, in-house financing is the financing scheme commonly offered as the last resort for home buyers, as this is the financing that has the highest interest rate. The advantage of in-house financing though is that it requires less paperwork and background check as opposed to Pag-IBIG financing and bank financing. Though the biggest disadvantage of in-house financing is that buyers have to pay a higher interest rate and as a result, have a higher chance of defaulting in their home loan equity.

Being disapproved in their home loan usually leads to a forfeiture of the house. This becomes a disadvantage for both the developer and the homebuyer, given that the developer shells out money for construction and documentary fees, while the buyer loses money on the down payment that has already been given. This particular example is one of the reasons why screening a buyer for his or her loan application is important even before the loan application documents are submitted to the selected financing institution. Loan application requirements that are usually asked from the homebuyers are the down payment of 10% to 20% of the contract price, employment documents, and post-dated checks to be submitted to the in-house financing institution. Interest rates of in-house financing are usually in between 14% to 18% with a loan term of either 5 or 10 years. Using Pro-Friends as an example, the interest rates of in-house financing are 18% or 21% with a loan payment term of 5 or 10 years, respectively. This type of financing usually has a bigger down payment for equity payment.

3. Predictive Modeling in Loan Applications

3.1. Binary logistic regression

Binary logistic regression has been used in the study of analyzing the behavior of credit and savings of cooperative societies in Rwanda (Papias and Ganesan, 2009). Papias and Ganesan (2009) found out that the Rwandan rural credit market is inefficient due to the poor credit repayment. They thus examined the factors that contribute to credit repayment behavior among members of savings and credit cooperatives in Rwanda. Different variables were used to come up with a logistic regression model for credit repayment rate. From the data they collected, Papias and Ganesan (2009) found out that variables such as age, gender, size of household, purpose for credit, interest rate charges, and number of official visits to credit societies are statistically significant in predicting credit repayment at 0.05 level. Variables such as size of credit disbursed, credit processing and disbursing time, borrowers' market place and income transfer from relatives and friends are significant at 0.20 level. All other variables, such as income, were found to be not statistically significant but have logical and/or explainable directions (Papias and Ganesan, 2009).

While there are different data mining predictive methods in classifying credit card applicants, one of the basic and easiest to appreciate is the logistic regression model (Wah and Ibrahim, 2011). Wah and Ibrahim (2011) suggest that demographic variables such as age, gender, and marital status are important to be included in credit scoring models because these variables can help classify which demographic is likely to have bad or good scores. Using the data that have been given to them by the Credit Card Center Department of banks in Malaysia, they were able to illustrate that older women are less likely to have a bad credit than young men. To which they have also found out that risk of default in loans decreases with age while risk becomes higher when applicants have dependents. Significant predictors in the logistic regression model are the following: age, gender, employment, property, number of loans, housing, home phone, duration, and loan history. The authors also compared logistic regression modelling to two other prediction procedures, CART and Neural Network (Wah and Ibrahim, 2011), and looked at the corresponding accuracy rates (see Table 4).

Table 4. Accuracy Rate of Predictive Models used by Wah and Ibrahim (2011)

Model	Training (75%)	Validation (25%)
Logistic Regression	72.53%	74.56%
CART	72.72%	73.66%
Neural Network	76.58%	76.46%

Credit scoring has been explored by Bensic, Sarlija and Zekic-Susac (2005) where they used binary logistic regression to separate the bad and good applicants. This can be applied in the real estate sense by replacing bad applicants with those that are likely to get their loan disapproved and good applicants with those that will likely have their loans approved. Bensic et al. (2005) used various prediction procedures including backprop neural network, RBFN, probabilistic NN, LVQ NN, logistic regression, and CART decision trees to create credit-scoring models on a sample of 160 applicants from a Croatian commercial bank (Bensic et al., 2005). The total hit rate which pertains to the accuracy of the model, hit rate of bads or true negatives relates to the proportion of bad applicants correctly predicted by the model, and hit rate of goods or true positives as the proportion of good applicants correctly predicted by the model. The final models of Bensic et al. (2005) had the following hit rates (see Table 5). It can be seen that logistic regression shows promise as compared to the other models as it had the second highest total hit rate, only behind Probabilistic NN.

Table 5. Hit Rates of Models used by Bensic et al. (2009)

Model	Total hit rate (%) or Accuracy	Hit rate of bad (%) or True Negatives	Hit rate of good (%) or True Positives
Logistic	76.32	53.58	88.00
Backprop NN	71.05	23.08	96.00
RBFN	71.05	76.92	68.00
Probabilistic NN	78.95	84.62	76.00
LVQ NN	65.79	15.39	92.00
CART	50.00	100.00	24.00

3.2. CART

In a particular research on prediction of default in loans, it has been found that CART can provide the best prediction with an average 8.31% error rate as compared to probit, neural networks, and k-nearest neighbors (Galindo and Tamayo, 2000). The different predictive models have been applied by Galindo and Tamayo (2000) in a mortgage loan dataset from a large commercial bank in Mexico provided by the *Comision Nacional Bancaraia y de Valores* (CNBV). A total of 4,000 mortgage loans from a single financial institution were used as sample and were analyzed. The single financial institution's mortgage loan portfolio represents roughly 14.3% of the total market in Mexico. The summary of the best models in the study can be seen in Table 6.

Table 6. Summary of Best Models by Galindo and Tamayo (2000)

Model	Test Error	Noise/Bias	Complexity
CART (120 nodes)	8.3%	0.073	21.7
Neural Net (16, 80)	11.0%	0.102	18.1
k-NN	14.95%	-	-
Probit	15.13%	0.150	1.80

Credit risk remains to be one of the most important variables in bank management, it can be related to the real estates' probability of loan disapproval from homebuyers. A study led by Eletter and Yaseen (2017) used three different decision models that discriminates "good" applications from "bad" risk applications. The models that were used were linear discriminant analysis (LDA), multi-layer perception (MLP) and CART decision trees. The data that were analyzed are from Jordanian Commercial Banks where a sample of 492 cases were collected. While the LDA model had the highest average correct classification rate and estimated misclassification cost (see Table 7), CART has still shown a promising result (Eletter and Yaseen, 2017).

Table 7. Accuracy Rate for Developed Models by Eletter and Yaseen (2017)

Model	Hit Rate of Good (%)	Hit Rate of Bad (%)	Total Hit Rate or Accuracy Rate (%)
LDA	85.8	70.5	79.2
MLP	94.0	91.8	93.1
CART	90.8	79.4	85.4

3.3. CTree and CHAID

CTree has been applied in exploring the risk and returns of investment in peer-to-peer lending (P2P). P2P is an emerging industry where members lend and borrow money using an online platform without intermediaries. Predictors that were explored by Singh et al. (2008), are credit grade of the client, purpose of loan, amount of loan, age, address, and employment status. Results show that credit grade is the most useful variable in analyzing the risk and predicting return on investment in peer-to-peer lending (Singh et al., 2008).

Similar to CTree, CHAID has been used in peer-to-peer lending. A study by Jin and Zhu (2015) used dataset covering 34,406 loan applications in Lending Club from the year 2007 to 2011 (details may be found in www.lendingclub.com). This study compared models of CART, CHAID, Multilayer Perceptron (MLP), Radial Basis Function (RBF), and Support Vector Machine (SVM). The results showed that SVM had achieved best performance in terms of accuracy in predicting default risk but the researchers said that improvement is very trivial over CHAID, the accuracy of which was closer to CART and MLP.

3.4. Random forest

Random Forest (RF) has also been applied in the world of business analytics by predicting credit risk of clients' in funding institutions. Machine learning approaches are usually more interpretable and thus it is easier to evaluate and explain the effect of each predictor in the prediction. Random Forest is said to perform better than most predictive models (Ghatasheh, 2014). Ghatasheh (2014) has used data from a German Credit dataset from the University of California in which there are 1,000 instances divided into 700 "good credit" and 300 "bad credit".

Ghatasheh (2014) compared five different random forest models which are Random Decision Forests (RDF), Random Forest, Random Forest Adaboost, Bagging RF, and C4.5 to predict credit risk (i.e. good or bad risk). The best models, Random Decision Forests and Bagging RF, produced a 78.4% accuracy which is better than the default C4.5 with an accuracy of 73.9%. He concluded that RF is competitive in terms of accuracy and simplicity and the randomness of RF produces better classification results.

The industry of peer-to-peer (P2P) lending also has research on methods that use RF to identify risk of a client. Results from a study by Malekipirbazar and Aksakalli (2015) showed that RF-based method outperforms FICO credit scores. FICO credit score is a score developed by Fair, Isaac, and Company to predict a consumer's ability to repay a debt on time. Its algorithm is proprietary to the FICO and it is the model used by a majority of American Banks.

In relation to the study by Malekipirbazar and Aksakali (2015), a team led by Kumar et al. (2016) also used RF in a P2P dataset. The study used Decision Tree, Random Forest, Extra Trees, and Bagging in predicting credit risk with RF having the highest accuracy. Kumar et al. (2016) found that RF has 88.5% accuracy which is 7.2% higher than Decision Trees. It is also noteworthy to include that while RF has a better accuracy in identifying loan defaults, Decision Tree is better in finding good credits.

4. Data and Methods

4.1. Data

The data collected is from a Philippine real estate company, Property Company of Friends, more commonly known as Pro-Friends. The data consist of 10,069 actual records of homebuyers from 2014 to 2016 who bought housing units from the developer. However, 753 buyers have been removed in the initial dataset since these buyers paid for their housing purchase in cash. These are the buyers who did not use a financing institution to pay for their home purchase. Summary of the number of loan applications for Pro-Friends Housing is shown in Table 8.

Table 8. Loan Applications for Pro-Friends Housing and Approved Housing Loans

Year	Number of Loan Applications	No. of Disapproved Housing Loans from Pag-IBIG and Banks	% of Disapproved Housing Loans from Pag-IBIG and Banks
2014	2,665	416	16%
2015	3,234	282	9%
2016	3,417	344	10%
TOTAL	9,316	1,042	12%

The number of clients who bought from Pro-Friends progressively increased from the year 2014 to 2016 while the number of disapproved housing loans in the three-year period is 1,042 out of the 9,316 clients. Characteristics of those who have gotten their housing loan approved and disapproved are explored in the predictive models utilized. The study focused on data that is available to provide an early and quick assessment of a clients' potential loan approval or disapproval. There were variables that were not explored in the study such as competition in the business, construction delays, etc., because of unavailability. There is also a number of clients who have forfeited their house purchase prior to their loan being approved. These clients have been removed from the dataset as Pro-Friends does not keep the profiles of forfeited accounts in compliance to the Philippine Data Privacy Act of 2012 (Congress of the Philippines, RA 10173, 2011).

The dependent variable, *loan application* is a nominal categorical variable. All the variables and their respective variable names, data type and definition can be seen in Table 9.

Table 9: List of Variables

Variable Name	Data Type	Definition
loan_application	Nominal	Whether loan is approved or disapproved by bank or HDMF
Sex	Nominal	Sex of the Buyer
Project	Nominal	Residential project name
Bedrooms	Discrete	No. of bedrooms of unit purchased
Bathrooms	Discrete	No. of bathrooms of unit purchased
Lot	Continuous	Lot area of house in sqm
Floor	Continuous	Floor area of house in sqm
lot_type	Nominal	Whether unit is inner, end, or corner
house_type	Nominal	Whether unit is duplex, townhouse, single attached, or single detached
Tcp	Continuous	Total contract price of unit in PHP
Finish	Nominal	Whether unit is bare or finished
Year	Ordinal	Year that the unit was reserved
employment	Nominal	Whether buyer is employed, self-employed, licensed professional, OFW, seafarer, and unemployed
civil_status	Nominal	Civil status of buyer
Age	Ordinal	Age of buyer in a range of years
income_buyer	Ordinal	Gross monthly income of buyer
reserve_fee	Continuous	Reservation fee of the unit purchased
Pdrf	Continuous	Promo given by developer on the purchased unit's reservation fee in PHP
Pdea	Continuous	Promo given by developer on the purchased unit's equity in PHP
payment_status	Nominal	Whether buyer is in-arrears or not after 3 months
document_status	Nominal	Whether documents are complete or incomplete after 3 months
equity_term	Discrete	No. of months needed to pay the equity or down payment

While this is actual data from Pro-Friends, it is noted that there are missing values in the dataset. Missing values may be attributed to wrong encoding or clients who do not want to disclose sensitive information to Pro-Friends. Only 1% of the data has missing values. Imputation will not be implemented in this research given that dataset only has a few missing observations. That is, cases/observations with missing values will be excluded in the analyses.

4.2. Methods

The dataset was divided into training and test sets. The training set was used to develop the model while the test set was used to address potential overfitting and/or to fine-tune the model. The dataset was divided into 75% training and 25% testing. The proportions of approved (89%) and disapproved (11%) housing loans have been maintained in the train-test splitting regardless of which year the loan applicant purchased a house.

Stepwise logistic regression is used in reducing the number of candidate variables in order to get the best performing model. AIC was used as the criterion for the basis of model reduction. For tree models, the pre-pruning criteria or "termination conditions" have been set via *maxdepth*, *minsplit,* and *minbucket* tuning parameters. *Maxdepth* is the tuning parameter used to set the maximum depth of the decision tree. Depth is defined as the longest length from the root node to a leaf node. Setting this parameter will stop the tree from growing when the set *maxdepth* is reached. *Minsplit* is the minimum number of records that must be reached in a node for a split to happen or be attempted. This would mean that if the minimum records in a split is set to be 20, then a node can be further split when the number of records in each split is at least 20. *Minbucket* is the minimum number of records that should be present in a terminal node. To illustrate, if the set minimum records in a node is 5, then the terminal/leaf nodes should have at least five records. Different runs have been made using these pre-pruning tuning parameters. Specifically, *minsplit* values were set at 5, 10, and 20, the *minbucket* were set at 5 and 30, and the *maxdepth* was set at 10. For random forest, the *mtry* parameter (i.e. the number of variables available for splitting at each tree node) has also been utilized to check if there are any improvements in the predictive power of the model. Specifically, *mtry* values of 3, 5 and 7 were explored.

The model performance is evaluated independently using the two subsets of training and test datasets. Performance of the models developed were compared in terms of the confusion matrices and hit rates (percentage correct classification). The candidate or best models used are the ones that have the high over-all hit ratio (or *accuracy*), *sensitivity* (correct classification among actual disapproval), *specificity* (correct classification among actual approval), *positive predictive value* or PPV (correct classification among predicted as disapproval), *negative predictive value* or NPV (correct classification among predicted as approval), and *F-score* (computed as twice the ratio of *PPV*sensitivity* and *PPV+sensitivity*). Higher values on these statistics suggest better predictive power of the models. Also, a higher F-score suggests balanced performance in predicting disapproval and approval of loans.

5. Results and Discussions

Table 10 refers to the summary statistics of quantitative (continuous and count) variables in the dataset. The average number of bedrooms that clients buy is 3. Bathrooms which comprises of toilet and bath in a housing unit has a mean average of 2, a minimum of 1, and a maximum of 2. Lot has a range of 28 sqm to 122.5 sqm with an average 61.20 sqm. Total contract price refers to the price of a housing unit which is in the P626,000 to 7.2 million range.

Males and females are roughly equal with 48% and 52% distribution respectively. LNC (or Lancaster New City) has the biggest share in the four projects of Pro-Friends with 86% while the next two biggest projects are Micara and Carmona Estates both with 7%. The most common lot type bought by clients is the inner lot with 77% of the customers choosing this over corner and end lots. This can also be attributed to the available inventory of residential housing as there are more. In terms of housing type, the townhouse is the most popular with 69% followed by Single Attached at 29%. Majority of the loan applicants have chosen a "bare" unit than a "finished" as it shows a split of 56% and 44% respectively.

Pro-Friends sales increases gradually from 2,665 in 2014, to 3,234 in 2015, and 3,417 in 2016. Majority of buyers are OFWs at 49% followed by locally employed clients at 47%. In terms of civil status, more are single at 53% than married at 46%. Majority of homebuyers are 30 to 39 years old and 40 to 49 years old, have monthly income between PhP 30,000 to 79,000, either employed or an OFW, and is not in-arrears. Only a few homebuyers are in-arrears in their equity payment (i.e. *payment_status*) and incomplete in their documentary requirements.

Table 10. Descriptive Statistics of Quantitative Variables

Variable	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Bedrooms	2	3	3	3	3	5
Bathrooms	1	2	2	2	2	3
Lot Area	40	50	75.00	54	100	308
Floor Area	28	50	61.20	60	70	122.5
Total Contract Price	623,160	1,344,000	1,833,666	1,534,450	2,187,000	7,231,680
Reservation Fee	0	10,000	11,927.92	10,000	20,000	30,000
Discount on Reservation Fee	0	0	1,039.93	0	0	12,500
Discount on Equity Amount	0	0	6,064.03	0	5,000	63,300
Equity Term	1	5	8	7	12	25

Looking at total contract price or TCP, majority of loan applicants have bought houses that are within the PHP 1.4M to 1.5M price range. This can be attributed to the fact that most houses offered by Pro-Friends are within the affordable segment. Reservation fee of most housing units are at PHP 10,000 followed by PHP 20,000.

Equity or down payment is usually paid for in a year by buyers but it should be noted there are many buyers who chose the 5-month term. Those who chose the latter can be attributed to people buying houses that are Ready-for-occupancy (i.e. houses that are already built and can be moved into after equity payment and loan approval). The lot area of most of the houses bought are 50 sqm while most of the houses have a floor area of roughly 30 sqm. Lastly, the dependent variable of loan application has an approval rate of 89% and thus only 11% are disapproved.

5.1. Binomial logistic regression

Different logistic regression models were considered. Other than the full model and the stepwise approach, fitting combinations of variables were also considered. These different runs were conducted to come up with a recommended predictive model for and/or get insights on factors that may influence the likelihood to be disapproved in a loan.

When running the full model of the logistic regression, all variables except *lot* are significant. The full model, however, seems to suffer from multicollinearity issues and/or confounding effects, as signs/directions of the regression coefficients are counter-intuitive. For example, the coefficients for *bedrooms* and *bathrooms* are negative while the coefficient for *TCP* is positive. This may warrant us to consider removing *bedrooms* or *bathrooms* (or both). Also, *payment_status* was excluded, as the sign is inverse of what is expected and/or intuitive.

Using the StepAIC (backward) approach, *lot type* and *bathrooms* were removed. The model includes variables *pr*oject, *lot*, *floor*, *TCP*, *type of house*, *finish*, *employment*, *civil status*, *age*, *sex*, *income of the buyer*, *discount on reservation fee*, *discount on equity amount*, and *document status*. With a forward StepAIC approach, all variables except *bathrooms* entered into the model.

There is also some potential in running models with interaction terms based on the insights done in the exploratory data analyses and from the initial runs. Interactions that were considered in the modeling were *Income_Buyer*Sex, Income_Buyer*lot, lot*Age, lot*finish, floor*finish, Income_Buyer*Age, TCP*house_type, TCP*Age, TCP*Civil_Status, finish*Age, Income_Buyer* Employment, lot*Employment*, and *TCP*Employment*. Of these (different models with) interactions, only *lot*finish*, *TCP*house_type* and *TCP*Income_buyer* were found to be significant.

A final logistic regression model was identified, taking into consideration predictive ability as well as utility and interpretation of the model. The significant predictors to get disapproved (or approved) for a loan are payment status, document status, employment, civil status, sex, age, income buyer, number of bedrooms, number of baths, floor area, type of house, and finish, as well as the interactions of TCP among income of the buyer and type of house. Payment status is the most important predictor in getting to know if an applicant will be disapproved or not, since it is an indicator of not being able to pay for a loan in the future. Self-employed applicants will also get a harder time applying for a loan because of unstable income which financing institutions look for in applicants. Number of baths and bedrooms are indicators of loan disapproval perhaps because having more baths and beds usually lead to a higher TCP, and a higher TCP usually lead to a higher loanable amount from financing institution. It can also be inferred that being single has a harder time in applying for a loan because of being unable to provide an automatic co-borrower for a loan given that co-borrowers increase the chance of loan approval. Summary of predictive ability are in Table 11.

Table 11. Results for Logistic Regression (training and test data sets)

	Predicted (Training) Cut-off = 0.11		
Actual	Disapproved	Approved	Total
Disapproved	184	510	694
Approved	3	6581	6584
Total	187	7091	7278

	Predicted (Test) Cut-off = 0.11		
Actual	Disapproved	Approved	Total
Disapproved	41	129	170
Approved	0	1646	1646
Total	41	1775	1816

	Training	Test
Accuracy	92.95%	92.90%
Sensitivity	26.51%	24.12%
Specificity	99.95%	100.00%
Positive Predictive Value	98.40%	100.00%
Negative Predictive Value	92.81%	92.73%
F-score	41.77%	38.86%

5.2. Classification and Regression Trees (CART)

CART has been used with varying pruning method specifications of *minsplit* of 5, 10, and 20, and with a *minbucket* and *maxdepth* of 5 and 10, respectively. No changes in the models were found even when changing *minsplit* and maxdepth to 25, 30, 35, and 40. However, changing *minbucket* to 30 removed one predictor in the decision tree. Results comparing the first CART model (*minsplit* = 5, 10, or

15*; minbucket* = 5*; maxdepth* = 10) versus the second generated model (*minsplit* = 5; *minbucket* = 30; *maxdepth* = 10) can be seen in Figures 1and 2.

The most important predictors in the first CART model (CART Model 1) are payment status followed by equity term, TCP, and *Project*. The second CART model (CART Model 2, see Figure 2) removed *Project* as one of the predictors as compared to the first one. The difference comes in the 6th level of the previous CART model (as presented in Figure 1) where the TCP of less than P1.5 million has been further split using *Project*.

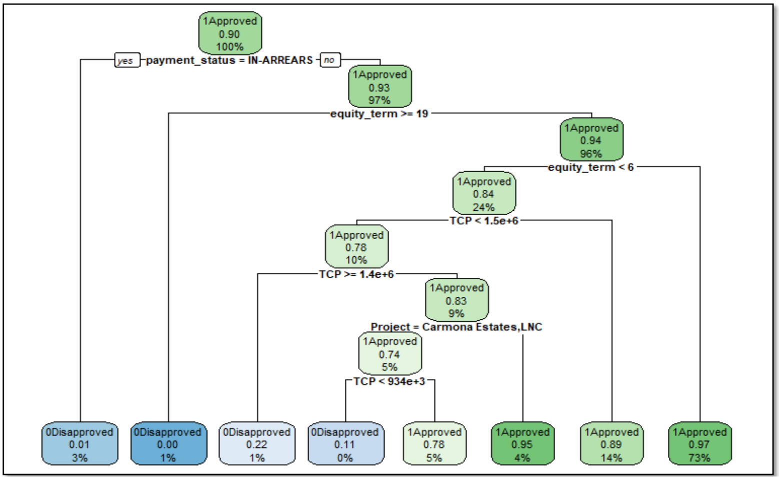

Figure 1. CART (Minsplit = 5, 10, 20; Minbucket = 5; Maxdepth = 10)

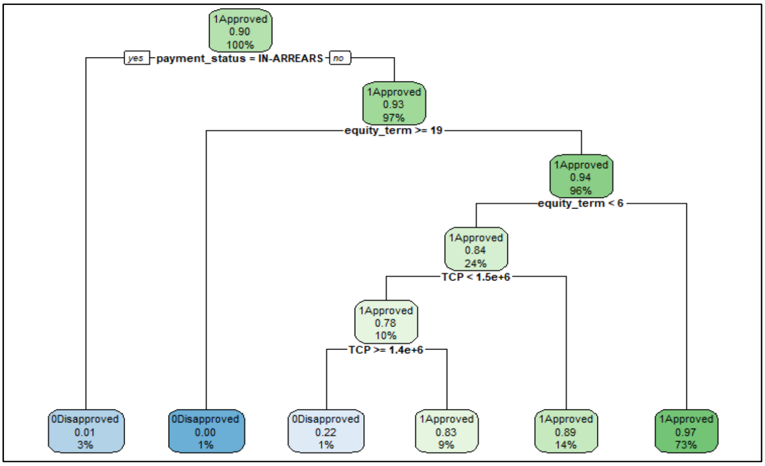

Figure 2. CART (Minsplit = 5, Minbucket = 30, Maxdepth = 10)

In terms of predictive power, both have a high accuracy rate with the CART Model 1 having an accuracy of 94.44% while CART Model 2 having an accuracy of 94.63%. While CART Model 2 is simpler in terms of number of nodes, CART Model 1 shines in having more variables used which in turn providing more insights for both developers and loan applicants. The confusion matrix results of CART Model 1 are in Table 12. Given the results, CART Model 1 is chosen not only because of performance but also because of the insights that it can provide.

Table 12. Results for CART Model 1(training and test data sets)

Actual	**Predicted (Training)**				**Total**
	Disapproved	**Row %**	**Approved**	**Row %**	
Disapproved	321	46.25%	373	53.75%	694
Approved	18	0.27%	6566	99.73%	6584
Total	339	4.66%	6939	95.34%	7278

Actual	**Predicted (Test)**				**Total**
	Disapproved	**Row %**	**Approved**	**Row %**	
Disapproved	40	23.53%	130	76.47%	170
Approved	0	0.00%	1646	100.00%	1646
Total	40	2.20%	1776	97.80%	1816

	Training	**Test**
Accuracy	94.44%	92.84%
Sensitivity	46.25%	23.53%
Specificity	99.73%	100%
Positive Predictive Value	95.02%	100%
Negative Predictive Value	94.41%	92.68%
F-score	60.10%	38.10%

5.3. Conditional Trees (CTree)

CTree provides more nodes and splits than CART which suggest different and various insights about home loan application disapproval. The predictors that CTree used are payment status, age, discount on equity, civil status, sex, equity term, finish, income of the buyer, and TCP (see Figure 3). Comparable to the initial split of CART, CTree also puts *payment_status* as the best predictor in determining loan disapproval. Being behind in equity payment would already qualify the applicant as disapproved. After *payment_status* the next split is on age, split into below 39 years old and 40 years old and above. Those that are younger (i.e. 39 years and below) and received a promo equity discount of more than PHP 10,000 are likely to get approved while those received a discount of less than PHP 10,000 will be further split into their sex. This can be attributed to the fact that discounts matter to the younger generation (i.e. millennials) than those who are older.

Although sex will be further split, it can be seen in the CTree plot that females are more likely to get disapproved in their loan application. In addition to this, males that do not receive a discount on their equity are more likely to have their home loan application disapproved. This does not support the research on predictive models on credit scoring where females are more likely to have good credit than bad.

Confusion matrices for the CTree model for training and test data sets are in Table 13.

Table 13. Results for CTree (training and test data sets)

Actual	**Predicted (Training)**				**Total**
	Disapproved	**Row %**	**Approved**	**Row %**	
Disapproved	218	31.41%	476	68.59%	694
Approved	16	0.24%	6568	99.76%	6584
Total	234	3.22%	7044	96.78%	7278

Actual	**Predicted (Test)**				**Total**
	Disapproved	**Row %**	**Approved**	**Row %**	
Disapproved	49	28.82%	121	71.18%	170
Approved	4	0.24%	1642	99.76%	1646
Total	53	2.92%	1763	97.08%	1816

	Training	**Test**
Accuracy	93.24%	93.12%
Sensitivity	31.41%	28.82%
Specificity	99.76%	99.76%
Positive Predictive Value	93.16%	92.45%
Negative Predictive Value	93.24%	93.14%
F-score	46.98%	43.95%

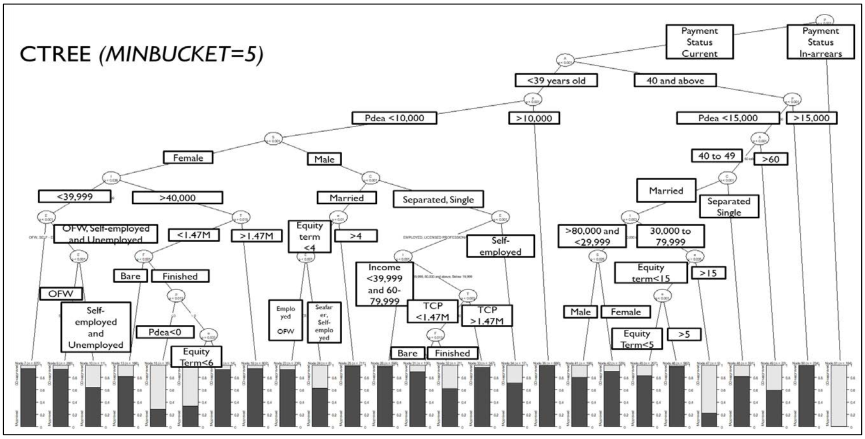

Figure 3. CTree Results (Minsplit = 5/10/20, Minbucket = 5, Maxdepth = 10)

5.4. CHAID

CHAID was ran three times with different parameters. Results show that there no significant changes when using *minbucket* = 5, 10, or 20. Accuracy of CHAID using different *minbucket* parameters remain above the 92% level. So instead of changing *minbucket,* the parameter *maxdepth* has been changed to make the model simpler and to check if accuracy has improved (or otherwise). The final CHAID decision tree can be seen on Figure 4.

Like the two previous decision trees, the most important predictor of loan disapproval is *payment_status*. This emphasizes the fact that not being updated in payment is an indicator that loan payment will be missed in the future. The difference of CHAID from CART and CTree is that splits are not anymore binary. It can be seen in the next node after *payment_status* is *age*, which is split into four different age groups – 18 to 39, 40 to 49, 50 to 59, and 60 and above. Those aged 40 to 49 have the highest chance of being disapproved for a loan especially when they are separated. This can be attributed to cases wherein loan application has been applied with a co-borrower, which is usually the spouse, and then separated during application period which led to a loan disapproval. Similar to CTree, CHAID also shows that self-employed and unemployed loan applicants would have a harder time in getting a loan approved because of inability to provide for a proof of stable income. Also similar to CTree, females are more likely to have their loan disapproved than males. The new variable that was introduced in CHAID, that was not present in CTree nor CART, is *FA* which shows that middle tier floor area (50 to 99 sqm versus <50 and >100 sqm) would have a higher chance of disapproval. This can compare to CART's middle-tier TCPs and CTree's middle-tier income being likely to get disapproved than those who are in the extremes.

The confusion matrix and evaluation measures for the CHAID model are summarized in Table 14.

Table 14. CHAID Results

Actual	**Predicted (Training)**				**Total**
	Disapproved	**Row %**	**Approved**	**Row %**	
Disapproved	182	26.22%	512	73.78%	694
Approved	2	0.03%	6582	99.97%	6584
Total	184	2.53%	7094	97.47%	7278

Actual	**Predicted (Test)**				**Total**
	Disapproved	**Row %**	**Approved**	**Row %**	
Disapproved	40	23.53%	130	76.47%	170
Approved	0	0.00%	1646	100%	1646
Total	40	2.20%	1776	97.80%	1816

	Training	**Test**
Accuracy	92.94%	92.84%
Sensitivity	26.22%	23.53%
Specificity	99.97%	100%
Positive Predictive Value	98.91%	100%
Negative Predictive Value	92.78%	92.68%
F-score	41.46%	38.10%

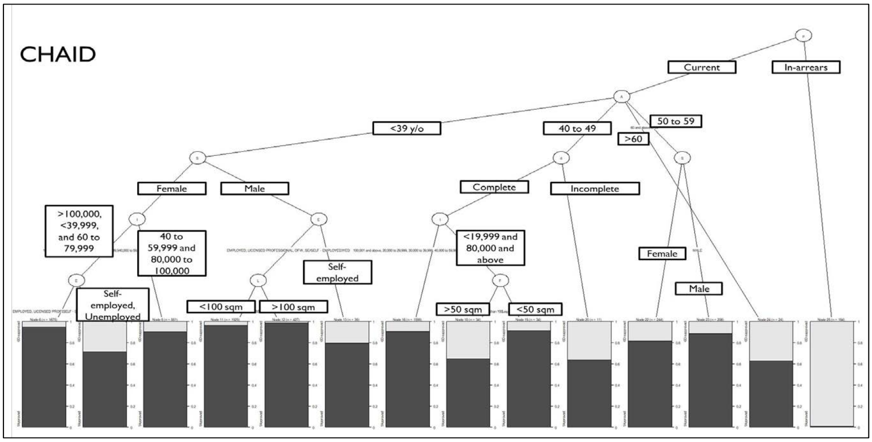

Figure 4. CHAID Results (minbucket = 5, minsplit = 5, maxdepth = 10)

5.5. Random Forest

Different random forest models were initially run, having varying number of trees (500 or 1000) and *mtry* at either 3 or 5. The model with 500 trees and *mtry = 5* has been chosen because it has a higher accuracy among the different models. The utilization of Random Forest in the dataset shows that the most important predictors in loan disapproval are TCP, payment status, and equity term, evident from their respective values of Mean Decrease Gini (see Figure 5).

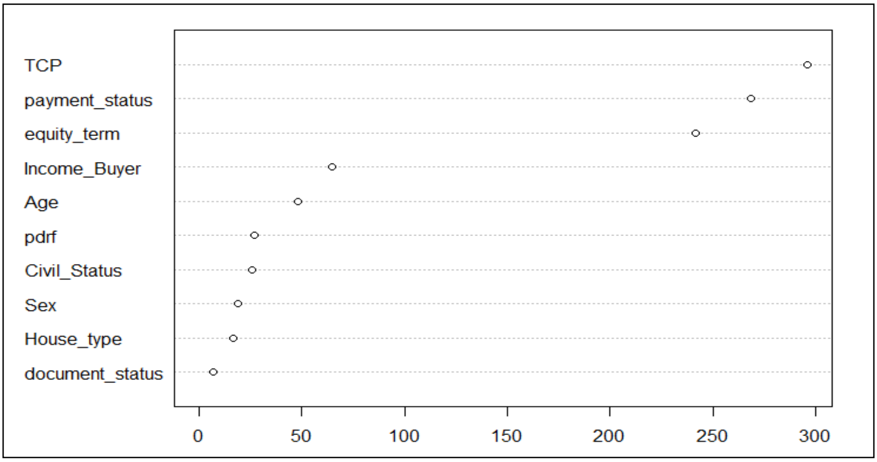

Figure 5. Random Forest Mean Decrease Gini

Results of the random forest show how important TCP, payment status, and equity term are in predicting loan disapproval. TCP should be commensurate to the income and equity term of the applicant (i.e. high TCP or more expensive housing units must have high income or if not, a longer equity term). Relative to the insights in other models, there will be higher probability of disapproval if the loan applicant purchased a house with a high TCP, is in-arrears with his or her payment, and has short-equity term. Payment status which is also present in all the models generated (i.e. CART, CHAID, CTree, and logistic regression) demonstrates the importance of not being in-arrears when applying for a loan.

The confusion matrices for the Random Forest model are shown in Table 15. The model has an accuracy of nearly 100% using the test data, the highest among all the predictive models used. Much better, with such a high accuracy level, Random Forest also maintained a high sensitivity score, at about 99% (test data), which is also way better than any of the decision trees or of the logistic regression model.

Table 15. Results of Random Forest

Actual	Predicted (Training)				Total
	Disapproved	Row %	Approved	Row %	
Disapproved	684	98.56%	10	1.44%	694
Approved	0	0.00%	6584	100.00%	6584
Total	684	9.40%	6594	90.60%	7278

Actual	Predicted (Test)				Total
	Disapproved	Row %	Approved	Row %	
Disapproved	149	87.65%	21	12.35%	170
Approved	3	0.18%	1643	99.82%	1646
Total	152	8.37%	1664	91.63%	1816

	Training	Test
Accuracy	99.86%	98.68%
Sensitivity	98.56%	87.65%
Specificity	100%	99.82%
Positive Predictive Value	100%	98.03%
Negative Predictive Value	99.85%	98.74%
F-score	99.27%	92.55%

Summary of Results using Test Data

The table below presents the different evaluation measures form the different predictive models using the test data on home loan disapproval. Among all the predictive models, the Random Forest yield the highest prediction accuracy, at almost 100%. Aside from overall classification power, it is evident that using Random Forest provides confidence in correct classification among the disapproval group, i.e., with Random Forest having the highest sensitivity or recall value (nearly four times than that of either logistic regression, CART or CHAID; triple than that of CTree). Specificity scores of the different models are comparable. Clearly, Random Forest has the most balanced predictive capability among all models (with Random Forest having the highest F-score).

Table 16. Summary of Confusion Matrices using Test Data

	Accuracy	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	F-score
Logistic	92.90%	24.12%	100.00%	100.00%	92.73%	38.86%
CART	92.84%	23.53%	100.00%	100.00%	92.68%	38.10%
CTree	93.12%	28.82%	99.76%	92.45%	93.14%	43.95%
CHAID	92.84%	23.53%	100%	100%	92.68%	38.10%
Random Forest	98.68%	87.65%	99.82%	98.03%	98.74%	92.55%

6. Conclusions

Different models are found useful in predicting home loan disapproval. Although logistic regression may already be appropriate to infer about possible factors that influence the likelihood of disapproval in a home loan, using actual data, decision trees and random forest are found to be beneficial and/or provide added value. CART, CTree, and CHAID highlight the potential non-linearity of the effects on likelihood to be disapproved (which, in contrast, is assumed to be linear in form under logistic regression). Random forest allows a more heuristic approach in identifying the factors and/or in predicting probability of home loan disapproval.

All the models implemented in the data on housing loans are competitive in that they all have high predictive capabilities, having accuracy rates of more than 90% (using test data). All in all, Random Forest is the best performing model and may be used for prediction given new data. The recommendation of using Random Forest is much supported being that model with the highest sensitivity and specificity, and so, over-all balance in accuracy of predicting disapproval or non-disapproval of housing loans.

The best variables in predicting loan disapproval are equity term, total contract price, payment status, and income of buyer. Several conclusions on predictors on loan application can be made from the models generated and based on the data used. From the results of the different models (and with focus on the recommended Random Forest), the following can be inferred:

The longer the equity term, the more likely the client will be approved for a home loan from Pag-IBIG and banks. This also means that the shorter the equity term, the more likely the client will be disapproved for a home loan.

Higher income of the buyer does not necessarily mean that a home loan will be approved – income must be commensurate to the TCP and equity term in the loan application. Although earning PHP 80,000 and above improves chances of loan approval.

A higher TCP on the other hand does not necessarily mean higher probability of disapproval. Those within the PHP1.4 to 1.5 million TCP are less likely to get disapproved among those in the lower income segment (less than P1.0 million). This may be attributed to the fact that Pag-IBIG Financing, which has longer payment terms, is targeted to the lower income.

Payment status, which is seen as one of the most important predictors across all models, shows that being behind in payment (i.e. in-arrears) would most likely lead to a disapproved loan.

Self-employed and unemployed applicants will likely have a harder time having a loan approved, potentially also because of unstable income.

In this dataset, females are more likely to get disapproved for a loan than males as compared to the research published by Wah and Ibrahim (2011), a study that was conducted using a dataset from Malaysia.

From the modeling exercise, it is recommended that the random forest model be used for predicting the likelihood to be disapproved of a housing loan. Being nonparametric and nonlinear, the recommended random forest model cannot be written nor presented in a structural function (or equation form). But, with the runs and/or implementation made in R, prediction of new data is possible with the saved random forest model object (i.e., a file stored in the R environment). Researchers and/or industry practitioners who may want to study further and/or use the recommended model may request for the said R model object directly from the authors (or respective institution).

Researchers and/or practitioners are also cautioned in the implementation and/or interpretation of the predictive model when used in their own or new datasets. As the recommended random forest model was derived from actual data, it is further suggested that the model be subject to regular reliability check (model maintenance and/or updating procedures). Test evaluation measures are relatively high for the random forest model, but the extent of utility/validity of the model is not explored (as it was not within the scope of this study).

References

ALONZO, A., 1994, The development of housing finance for the urban poor in the Philippines: The experience of the Home Development Mutual Fund. *Cities*, pp. 398-401.

BALLESTEROS, M. M., 2010, December, Linking Poverty and the Environment: Evidence from Slums in Philippine Cities, *PIDS Discussion Paper, 33*, pp. 1-32.

BANCO DE ORO, 2019, March 1, *Various Home Loan Options*, Retrieved from Banco de Oro: https://www.bdo.com.ph/mobile/personal/loans/home-loan

BENSIC, M., SARLIJA, N., and ZEKIC-SUSAC, M. 2005, Modelling Small Business Credit Scoring by Using Logistic Regression, Neural Networks, and Decision Trees, *Intellectual Systems Accounting and Financial Management*, pp. 133-150.

BRIEMAN, L. (2001). Random forests, *Machine Learning*, pp. 5-32.

CONGRESS OF THE PHILIPPINES. (2008, July 28). *Republic Act No. 9679*. Retrieved from Senate of the Philippines: https://senate.gov.ph/republic_acts/ra%209679.pdf

___________, 2011, July 25, *RA 10173*. Retrieved from National Privacy Comission: https://www.privacy.gov.ph/data-privacy-act/

___________, 2017, December 13, *Republic Act No. 10963*. Retrieved from Senate of the Philippines : https://www.senate.gov.ph/republic_acts/ra%2010963.pdf

___________, 2019, February 14, *Republic Act No. 11201*. Retrieved from Official Gazette of the Philippines: https://www.officialgazette.gov.ph/downloads/2019/02feb/20190214-RA-11201-RRD.pdf

ELETTER, S., and YASEEN, S., 2017, Loan decision models for the Jordanian commercial banks, *Global Business and Economics Review*, pp. 323-338.

FERNANDO, A., LUCAS, R. E., and WEBER, E. W. (2001). Interest rates and inflation, *American Economic Review*, pp. 219-225.

FRANCO-GARCIA, K.-L. N., and GALANG, JR., F. O., 2017, *Pag-IBIG Fund's Interest for Affordable Housing Lowest at 3%.* Makati City: Pag-IBIG.

GALINDO, J., and TAMAYO, P., 2000, Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*, pp. 107-143.

GENERALAO, M., 2017, July 22, *Strategy to meet 12.30-M housing need*. Retrieved from Philippine Daily Inquirer Net: https://business.inquirer.net/233654/strategy-meet-12-3-m-housing-need.

GHATASHEH, N., 2014, Business analytics using random forest trees for credit risk prediction: A comparison study, *International Journal of Advanced Science and Technology*, 19-30.

HDMF, 2019, March 1, *Housing Loan Amortization Calculator*. Retrieved from Pag-IBIG Fund: https://www.pagibigfund.gov.ph/AA/calc.aspx.

HOUSING AND LAND USE REGULATORY BOARD, 2018, May 2, *HLURB Memorandum Circular No. 09*. Retrieved from Housing and Land Use Regulatory Board: http://hlurb.gov.ph/wp-content/uploads/memorandum-circulars/2018%20 MC/MC-18-09.pdf.

JIN, Y., and ZHU, Y., 2015, A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending. *Fifth International Conference on Communication Systems and Network Technologies*.

KUMAR, V. L., NATARAHAN, S., LAKSHMI, K. N., and LAKSHMI, C. N., 2016, Credit Risk Analysis in Peer-to-Peer Lending System, *IEEE International Conference on Knowledge Engineering and Applications*, pp. 193-196.

MALEKIPIRBAZARI, M., and AKSAKALLI, V., 2015, Risk assessment in social lending via random forests, *Expert Systems with Applications*, pp. 4621-4631.

PADOJINOG, W., VILLEGAS, B., TEROSA, C., HUBO, C. L., ALFORTE, J., JANEO, V., and CASTILLO, E., 2012, *The Housing Industry Road Map of the Philippines: 2012-2030.* Pasig: Subdivision ahd Housing Developers Association and Center for Research and Communication University of Asia and the Pacific.

PAPIAS, M., and GANESAN, P., 2009, Repayment behavior in credit and savings cooperative societies: Empirical and theoretical evidence from Rwanda, *International Journal of Social Economics*, pp. 608-625.

PATNAIK, B., SATPATHY, I., and SAMAL, N., 2017, Home Loan Portfolio - A Review of Literature, *International Journal of Current Advanced Research*, pp. 5804-5807.

PHILIPPINE STATISTICS AUTHORITY., 2017, June 30, *Philippine Population Surpassed the 100 Million Mark (Results from the 2015 Census of Population).* Retrieved from Philippine Statistics Authority: https://psa.gov.ph/population-and-housing/node/120080

_______ 2018, March 6, *Housing Characteristics in the Philippines: Results of the 2015 Census of Population.* Retrieved from Philippine Statistics Authority: https://psa.gov.ph/population-and-housing/node/129804

_______ 2019, February 28, *Inflation Rates (Philippines)*. Retrieved from Bangko Sentral ng Pilipinas: http://www.bsp.gov.ph/statistics/spei_new/tab34_inf.htm

RIVAS, R., 2019, January 5, *Buying a home in 2019? High interest rates will bite*. Retrieved from Rappler: https://www.rappler.com/business/220308-high-interest-rates-will-bite-home-loans-2019

TSATARONIS, K., and ZHU, H., 2004, What drives housing price dynamics: cross-country evidence. *BIS Quarterly Review*.

WAH, Y., and IBRAHIM, I. (2011). Using Data Mining Predictive Models to Classify Credit Card Applicants. *Expert Systems with Applications*, 13274-13283.