Applied Machine Learning for Central Bank Statistics: Supervised Models for The Detection Of Subsidized Housing Complexes In Chile

Martín Rebolledo Jaure *mrebolledo@bcentral.cl*
Statistics and Data Division
Central Bank of Chile
July, 2023

Abstract

Each trimester, the Central Bank of Chile calculates the Household Price Index. This publication serves as an indicator of the price trends of the country's housing market. The index is based on administrative records of housing transactions, which include various types of properties. To ensure the accuracy of the index, transactions that are part of fully subsidized social housing complexes must be identified and excluded. This paper addresses this need by proposing a supervised classification approach as a binary statistical classification problem using Machine Learning Models. The chosen model, a Random Forest Model with measures to prevent overfitting, achieves a high rate of recall and accuracy in detecting social housing transactions. Results show that from 2004 to 2022, 9% of all transactions in Chile corresponded to social properties. Furthermore, our study highlights the potential of machine learning for automating data processing in Central Banks, which can lead to enhanced accuracy in the creation of official statistics.[1]

Keywords: Big data, Automatization, Social Housing, Official Statistics, Machine Learning, Central banking.

JEL classification: C18, C53, C58, C81, E47, E58, G22.

[1]This paper was made in the context of the 64th ISI World Statistics Congress. The views and conclusions presented in this paper are exclusively those of the author and do not necessarily reflect the position of the Central Bank of Chile or of the Board members. This study is conducted within the research agenda developed by the Central Bank of Chile on economic and financial matters under its competences. Within this framework, the Central Bank has access to anonymized information from various public and private entities, based on collaboration agreements signed with these institutions. The information contained in the Internal Revenue Service (SII) databases is of a tax nature, stemming from self-declarations of taxpayers submitted to the SII. The accuracy of this data is not the responsibility of the Service. We thank comments and suggestions from Juan Pablo Cova, Juan José Balsa, Patricia Medrano, and Javiera Vásquez.

1 Introduction and Motivation

Central Banks are increasingly embracing Data Science applications on their operations, recognizing its transformative potential. Machine learning algorithms, predictive analytics, and advanced statistical models are among the primary tools being used for this purpose. They facilitate the processing and understanding of large data sets, enabling Central Banks to comprehend macroeconomic trends, evaluate financial stability, formulate robust monetary policies and make better official and experimental statistics.

This paper exposes one applied use case of Machine Learning models in the calculation of the Household Price Index (HPI) published by the Central Bank of Chile. The HPI serves as an indicator of the price trends of the national residential-housing market in Chile. It uses administrative data from effective housing transactions in the country from the Chilean Internal Revenue Service. As this index intends to evaluate the buyers and seller market, a big portion of the data preprocessing is dedicated to identifying those transactions that do not belong to the traditional market. One of those sources of anomalies is subsidized housing transactions.

Chile has a well-established policy regarding subsidized housing (Hidalgo, 2019). Approximately 42% of homeowners in the country acquired their properties by using government subsidies, with nearly 15% financing their entire property cost utilizing these subsidies (Central Bank of Chile, 2022). A notable aspect of this policy is the emphasis on giving complete ownership of the property to the families. Therefore, the transfer of subsidized housing units to the families is registered in the same manner as transactions within the buyer's and seller's property market.

In the Bank, the detection of projects of subsidized housing properties has been conducted since the beginning of the calculation of the index through manual revision by Central Bank analysts. That work made available a training data set of subsidized housing projects, that now allows us to automatize the process.

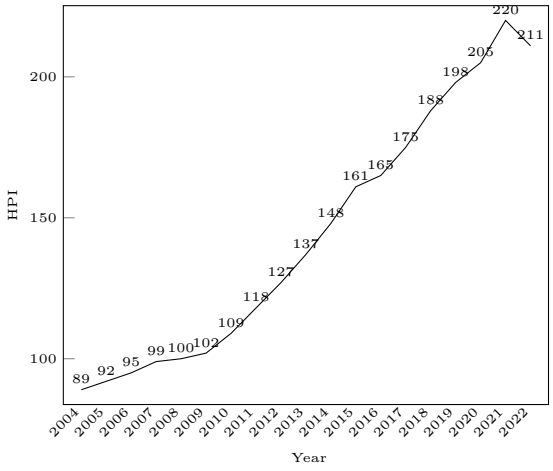
This paper proposes a ML approach to solve a data processing issue on the calculation of the Household Price Index. More in detail, as part of the necessary data management process, the Bank needs to detect and exclude social housing transactions from the data set. However, in the administrative data used for the calculation of the indicator, there are no variables that indicate that a transaction should be catalogued as social housing transaction.

Most of the literature on central bank applications is included in the Bulletin 57 of the Irving Fisher Committee (Araujo et al., 2022), which exposes many experiences of the use of Data Science in different Central Banks. Between other experiences, we highlight La Serra & Svezia (2022) who show the benefits of trying different ML supervised classification models in detecting anomalies in administrative records of insurance corporations' reports. Also, to Maddaloni et al. (2022) who go one step ahead, by proposing a stacked approach that combines multiple ML models for anomaly detection in the Ana credit database in Europe.

As for this paper, We propose a supervised classification approach that considers ML and statistical algorithms. We build up new variables from the administrative records and try four different ML models of binary classification: regularized weighted logistic regression, kernel weighted support vector machines, regularized neural network using dropout regularization methods and Weighted Random Forest.

The paper is organized as follows. Section 2 describes the Household Price Index and the motivation for this work. Section 3 introduces social projects in Chile and their characteristics. In section 4 we describe the data from which the analysis is derived. In section 5 we discuss the models used. Section 6 shows the main results of the Social Project classification and the discussion about the model to choose. Section 7 exposes the main results for subsidized housing and section 8 exposes the operability of this solution as a usual process of the Central Bank of Chile. Section 9 summarizes the main conclusions, showing the benefits of this work for applied ML in the Central Banking community.

Figure 1: Historical Serie of the Household Price Index
(2008=100)



Retrieved from *Statistics Database*, Central Bank of Chile, 2023

2 The Household Price Index

Since 2014, the Central Bank of Chile computes the Housing Price Index (HPI). This indicator is derived from administrative microdata provided by the Internal Revenue Service (SII), which includes actual transactions of residential properties registered in the F2890 Form of "alienation and registration of properties". This information is further supplemented with data from the Real Estate Registry (CBR). In Figure 1 we can see the HPI index overtime. The year base is 2008 (2008=100).

The HPI serves as a critical tool to analyze the real estate sector and understand the dynamics of housing prices in Chile. Housing stands as the preeminent asset for households and constitutes a major component of their expenditures. Changes in housing prices directly impact households' wealth and borrowing capacity, thus affecting consumer spending. Additionally, any instability in the housing market could pose risks to the financial health of lending institutions, potentially triggering a wider financial crisis.

At present, the HPI is published quarterly and consists of nineteen series: the general HPI, broken down into houses, apartments, new and used, and for seven geographical zones, dating back to the first quarter of 2002. The HPI is published in the "Experimental Statistics" and "Sectoral Statistics" chapters of the Statistical Database on the Central Bank of Chile's website.

Its computation involves a mixed stratification methodology. This calculation entails numerous stages of data cleansing and imputation of significant variables. For a more detailed understanding of the index's calculation methodology, refer to Flores & Pérez (2015) and to Balsa & Vásquez (2023).

3 Social Housing in Chile

Chile has a public policy on social housing. In 1959, the nation introduced the National Housing Program (Programa Nacional de Vivienda, in Spanish), which has fundamentally maintained its essence over the years. According to Salvi Del Pero and colleagues (2016), the primary objective of this policy is to mitigate the housing shortage faced by vulnerable populations, while also assisting the middle class in fulfilling their housing aspirations.

In Chile, 42% of households that own a property used a subsidy for its acquisition. Meanwhile, 14.5% of properties were entirely financed through a subsidy (Central Bank of Chile, 2022). These statistics underscore the significant role of government subsidies in facilitating homeownership in the country, highlighting how these measures have enabled a significant proportion of the population to secure housing.

The policy framework of subsidized housing in Chile stands out for its emphasis on endowing beneficiaries with full ownership of the given properties. Unlike other models, such as those prevalent in European countries like the United Kingdom, Netherlands, and Sweden, where subsidized or social housing is typically owned by public or non-profit entities and provided to beneficiaries under rental or lease agreements, Chile's policy hands over the property rights to the beneficiaries.

The government extends a variety of subsidy alternatives encompassing aspects such as property acquisition, rental assistance, and home repair subsidies, to name a few. However, for the purposes of this paper, our attention will be specifically directed towards identifying properties that are entirely subsidized and belong to a newly established neighborhood of households. The next table specifies the kind of subsidies offered by the government to these ends [2].

Table 1: Comparison of Housing Subsidies in Chile

Subsidy Program	**Eligibility**	**Benefit**	**Requirements**
Solidarity Fund for Housing Choice (Fondo Solidario de Elección de Vivienda)	Designed for the poorest sectors of the population	Grants a subsidy to buy a new house or to improve or extend a current property	The beneficiary must have been saving consistently for a minimum period in a Housing Savings Account
DS1 Subsidy (Decreto Supremo No. 1)	Low-income families that cannot afford a home without financial support	Provides a subsidy towards the full or part of the payment of a property	The property must be under a certain value, and the applicant must have a certain number of savings in a Housing Savings Account
DS19 Subsidy (Decreto Supremo No. 19)	Low and Middle-class families who cannot afford a home without financial support	Provides a subsidy for the full or part of the payment of a property	The property must belong to specific approved project, and the applicant must have a certain number of savings in a Housing Savings Account

Elaborated by the authors using public information from the Chilean Ministry of Housing and Urban Planning

The three subsidy programs presented in the preceding table extends comprehensive financial support for property acquisition to the most vulnerable sections of the population. Beneficiaries have the flexibility to either apply their subsidy towards new social housing complexes or finance an existing property. For the scope of this study, we will narrow our focus to social housing complexes, a sector where property prices are significantly lower than those observed in traditional housing transactions.

Subsidized housing transactions represent a complication for the Housing Price Index (HPI) since the prices recorded for the handover of properties within subsidized housing complexes to beneficiaries are documented in the same way as transactions within the buyer's and seller's real estate market, making their exclusion

[2]Please note that the information outlined in this document serves as a general guide to the various housing subsidies currently available in Chile. It's worth mentioning that the precise financial details and qualifying criteria of these subsidies may differ from the specifics indicated here. As a result, it is vital for readers, especially those planning to avail of these subsidies, to engage with relevant Chilean government bodies like MINVU or other trustworthy resources to get the most up-to-date and complete data on housing subsidies.

complex. This circumstance contributes to an underestimation of the index. Moreover, the administrative records lack of indicators to differentiate between a standard property and a social housing unit. Consequently, this situation necessitates that we devise a solution that will enable us to impute this variable and subsequently exclude such properties from the calculation.

For the purpose of this paper, we define Subsidized Housing transactions as property transactions fully funded by the government that are granted to low-income families. Since the starting of the index's computation, the identification of subsidized housing projects has been carried out through manual revision by Central Bank analysts. This effort has resulted in the creation of a training data set for subsidized housing projects, which now enables us to automate the process.

4 Data Description

The Central Bank of Chile, in the context of their regular publication of national-wide statistics, receives data from many sources. In particular, the granular data presented in this study comes from the Chilean Internal Revenue Service (SII) sourced from the F2890 form, supplemented with information from the Real Estate Register (CBR). This data focuses on individual property transactions.

The data set contains valuable information of the characteristics of the transaction of properties. Key data points include the transaction price, the size of the construction in square meters, the size of the land of the property in square meters, the date of the transaction, a dummy variable that indicates if the buyer funded the property using a mortgage loan, a dummy that indicates if the property transfer was made by a company or a person, among other useful columns.

The original Data set comes in a property level. For this paper, we aggregate the data set in a block-of-properties level, which mean that for each data points we will be having information for a group of properties aggregated in blocks, which is a well defined geographical group of properties, constructed by the Internal Revenue Service. This implies that for the forthcoming models, we will be having predictions for the whole block, rather than for individual properties, which is coherent with the structure of a subsidized housing block of properties, that is the target of prediction.

Table 2 reveals substantial differences between blocks of non-subsidized and subsidized housing, suggesting clear differences between their characteristics. These disparities are potentially promising for our forthcoming analyses, as they may facilitate separability in the data, while enhancing the efficacy of our predictive models. This clear distinction between groups could be instrumental in creating robust models that can accurately capture and reflect the dynamics of social and regular housing sectors.

Table 2: Interest variables between subsidized and regular housing blocks

Average of Interest Variable	**Subsidized Housing**	**Regular Housing**
Property Price (Thousands of Dollars)	42.5	125.7
Construction Area in Square Meters (m^2)	51	70
% of Properties Purchased with Mortgage in the Block	1.6%	46.7%
% of Properties Sold by the Same Seller in the Block	91%	61%
% of Properties Sold by a Legal Entity in the Block	88%	36%

Elaborated by the authors using information from the Internal Revenue System

5 Supervised ML for classification

In supervised learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data by associating patterns to the unlabeled new data. For this paper we propose the use of three widely used models for binary prediction: First, regularized weighted logistic regression. Second, kernel weighted support vector machine. Third, regularized neural network using dropout regularization methods and fourth, Random Forest. We separate the data in train and test subdatasets.

5.1 Logistic Regression for classification (LR)

Logistic regression is a method widely used in economics. Considering their probabilistic approach it has many applications in Machine Learning. According to Dreiseitl and Ohno-Machado (2002) it provides a functional form f and parameter vector α to express $P(y \mid x)$ as $P(y \mid x) = f(x, \alpha)$. The parameters α is determined based on the data set D, usually by maximum-likelihood estimation with parameters can be interpreted (parametric method).

In the current data structure, approximately 8% of the blocks are social project blocks. For that, we implemented the suggestions of King and Zeng (2001) for imbalanced data. That is implemented in Python by the fixation of the weights by an hyperparameter grid search that allows us to maximize AUC and recall score. In addition, as for preventing the model to overfit we make use of ridge regression regularization methods implemented in the Skcikit-learn library in Python (Pedragosa et al., 2011).

5.2 Support Vector Machine

This learning strategy introduced by Vapnik 1998 is a principled and very powerful method that has outperformed most other models in a wide variety of applications (Cristianini & Scholkopg, 2002). Following Maldonado et. Al (Maldonado et al., 2011), Given training vectors $x_i \in R^n, i = 1, \ldots, m$ and a vector of labels $y \in \mathbb{R}^m, y_1 \in \{-1, +1\}$, SVM provides the optimal hyperplane $f\mathrm{x} = w^T x + b$ that separate the training patterns. If the case of linearly separable classes, this hyperplane maximizes the sum of the distances to the closest positive and negative training patterns.

For not separable cases, the literature suggest that the hyperplane can be found by using kernel functions to compute dot products in a higher-dimensional space and use those to find a hyperplane. According to Cristianini and Scholkopf (2002) the fundamental idea of kernel methods is to embed the data into a vector space, where linear algebra and geometry can be performed. One of the simplest operations one can perform in such spaces is to construct a linear separation between two classes of points. Other operations can involve clustering the data or organizing them in other ways. The use of linear machines is easier if the data are embedded in a high dimensional space.

We may argue that for the detection of social projects this is the case (non-separable). We can introduce that kernel-based solution in Python by using the Skcikit-learn library (Pedragosa et al., 2011). We use radial basis functions (RBF) for the kernel, which approximate any function under mild conditions (Benoudjit & Verleysen, 2003).

5.3 Artificial Neural Networks

According to Faris, Aljarah and Mirjalili (2016) Artificial Neural Networks (ANNs) are mathematical used for modeling complex nonlinear processes. Some of the attractive characteristics of ANNs include the ability to capture nonlinearity, they are highly parallel their fault/noise tolerance (Basheer & Hajmer, 2000). As for this paper, we make use the feedforward Multilayer Perceptron (MLP).

The model consists in 3 dense hidden layers with ReLu activation functions and a final output layer with sigmoid activation function to produce a binary classification prediction. Dropout regularization (Srivastava et al., 2014) is applied after each hidden layer to prevent overfitting and improve the generalization performance of the model. The model is trained using the binary cross-entropy loss function and the Adam optimization algorithm. The performance of the model is evaluated on a test set, and the result show that the proposed model achieves high accuracy and generalization performance on the binary classification task.

5.4 Random Forest

According to (Breiman, 2001), Random Forests (RFs) are an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Key attributes of RFs include their robustness to overfitting due to bootstrapping and bagging techniques, high scalability, and inherent feature selection due to the random subspace method (Cutler et al., 2012). For the purposes of this paper, we employ a random forest classifier.

Our chosen model consists of 100 decision trees, each constructed by randomly sampling from the feature and instance spaces, reducing correlation between trees and mitigating the possibility of overfitting. Out-of-bag samples are used for internal cross-validation during training to provide an unbiased estimate of the model's performance. The Gini impurity criterion is used to split nodes, enabling the model to handle categorical data effectively. Performance of the random forest classifier is evaluated on an independent test set. Results indicate that our model demonstrates high predictive accuracy and impressive generalization capabilities in the binary classification task.

6 Model results and selection

6.1 Model results

Results are displayed in the subsequent Table 3, including accuracy, recall, AUC, F1, and Matthews correlation coefficient for each of the four models. In short, the accuracy score quantifies overall correct predictions. Recall indicates the proportion of correctly identified positives, crucial for minimizing false negatives. AUC measures a model's ability to distinguish between classes; the closer to 1, the better. F1 score is the balance between precision and recall, key when both false positives and false negatives are significant. Lastly, Matthews correlation coefficient evaluates the quality of binary classifications, providing a balanced measure despite different class sizes.

The upcoming subsection will delve into the discussion of the results. We also provide a graphical representation of the Receiver Operating Characteristic curve (ROC). The ROC curve is a plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, displaying the trade-off between sensitivity (or true positive rate) and specificity (or false positive rate).

6.2 Discussion

While all four models demonstrate high accuracy scores, with Random Forest leading at 98.84%, accuracy is not the primary measure of performance due to the imbalance of the data. In this context, the recall score is particularly crucial because it reflects the model's ability to correctly identify true positives. For us, high recall is critical because false negatives would mean missing subsidized housing projects, which is a highly costly error. The Random Forest model outperforms the other models with a recall score of 84.87%, reducing the chance of such costly misses. The results shown in table 3 are constructed using the test data.

Table 3: Performance of supervised ML models

Model	SVM	ANN	Logistic	RForest
Accuracy Score	97.45%	98.74%	98.62%	98.84%
	(97.15%, 97.71%)	(98.47% , 98.94%)	(98.44% , 98.82%)	(98.66% , 99.01%)
Recall score	87.76%	77.08%	76.92%	84.87%
	(84.89%, 90.55%)	(72.87% , 81.02%)	(73.77% , 80.1%)	(81.67% , 88%)
Area Under Curve	92.83%	88.43%	88.28%	92.19%
	(91.41%, 94.24%)	(86.31% , 90.36%)	(86.73% , 89.86%)	(90.63% , 93.75%)
F1 Score	75.77%	84.81%	83.44%	86.98%
	(73.17%, 78.39%)	(81.84% , 87.07%)	(81.38% , 85.84%)	(84.93% , 89%)
Matthews Coeff.	75.24%	84.63%	83.06%	86.41%
	(72.62%, 77.82%)	(81.72% , 86.85%)	(80.97% , 85.51%)	(84.29% , 88.48%)

Elaborated by the authors using information from the Internal Revenue System

Figure 2: Receiver Operating Characteristic Curve (ROC) of the models

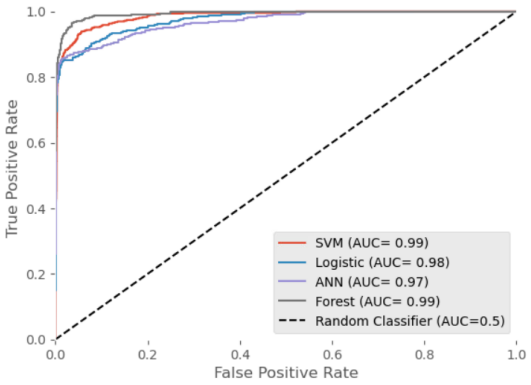
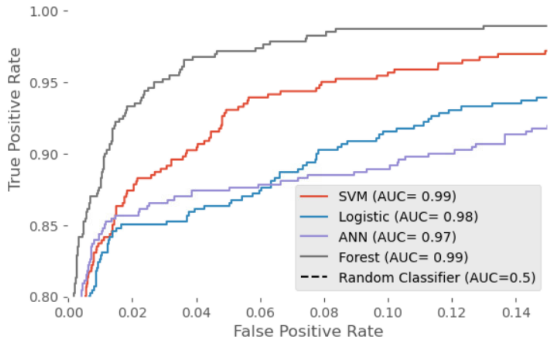


Figure 3: Zoomed on the Receiver Operating Characteristic
(ROC)



The F1 score and Matthews correlation coefficient, which provide a balanced measure of the model's performance, also favor Random Forest with values of 86.98% and 86.41% respectively. Meanwhile, the Area Under Curve (AUC), demonstrating the model's ability to differentiate between the classes, is the highest in the Random Forest model at 92.19%. This indicates the model's superior performance in class separability.

Although the ANN model shows similar accuracy as Random Forest, its lower recall score compared to Random Forest indicates less optimal performance in correctly identifying true positives. This is a critical consideration given the high costs associated with false negatives in our context.

Given the superior performance of the Random Forest model across these critical metrics, it emerges as the most suitable choice for this binary classification task, despite the data imbalance. It aligns with our priority of minimizing false negatives and is thereby the most effective tool for predicting subsidized housing projects. This is coherent with the literature in this respect. For a few reasons:

Primarily, Random Forest uses an ensemble of decision trees that vote collectively to produce a final prediction. This technique offers several advantages: it counteracts the overfitting problem common in single decision trees and enhances the model's predictive power.

Specifically for imbalanced datasets, as in our case, Random Forest is an effective choice. This model's ensemble nature allows it to better identify and classify observations from the minority class, even when that class represents a small fraction of the total observations, as with the treatment class in our data. This strength emerges because each individual decision tree in the forest gets a fair chance of being trained on different data subsets, including various representations from the minority class.

Additionally, Random Forest doesn't require extensive data preprocessing and implicitly performs feature selection, both contributing factors to its performance superiority.

This results are coherent with the literature. Li Ying (2018) compared three families of models: random forest, logistic regression, and SVM on bank loan data. The results show that random forests generally perform better. This result corresponds with a study of Coser, Maer-matei Albu (2019)that compared LightGBM, XGBoost, Logistic Regression, random forest to evaluate loan probability default and recognized random forest as an optimal classifier for the task. The learning rate is fast and can be applied to large-scale datasets (Wang et al., 2020).

7 Social Housing Results

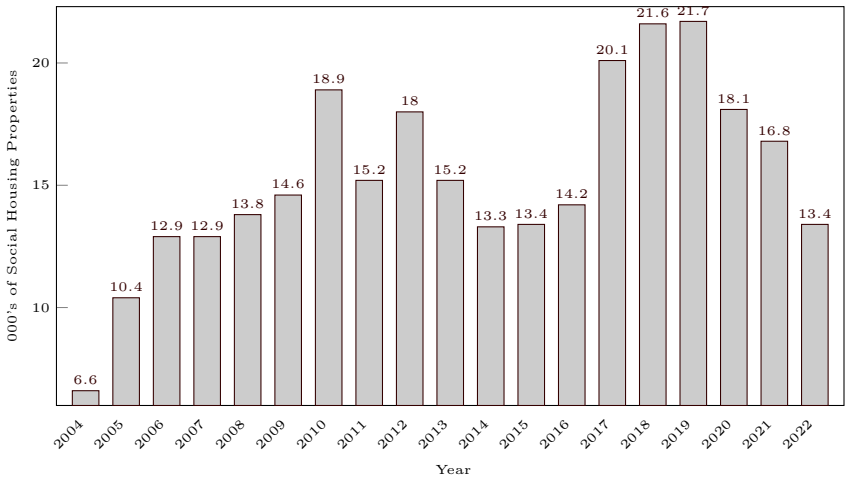
We estimate that, for the period of 2004 to 2022, nearly 300,000 properties are likely to be properties belonging to social projects blocks, which represent nearly an 9% of all transfer of properties in Chile. In terms of geographical distribution, the south of Chile concentrates most of the social households, a 15% of all transactions are likely to be social. Table 4 exposes that distribution. Figure 2, exposes the estimation of the number of subsidized housing units delivered by year.

Table 4: Distribution of Subsidized Housing by Geographical Zone (2004-2022)

Region	**N of Transactions**		**Subsidized (%)**
	Total	Subsidized	
North of Chile	242.210	21.573	9%
Center of Chile	855.301	99.449	12%
South of Chile	700.485	115.836	17%
Santiago	1.483.523	54.248	4%
Totals	**3.281.519**	**291.106**	**9%**

Elaborated by the authors using information from the Chilean Internal Revenue System

Figure 2: Number of Subsidized Housing by Year (2004-2022)
000's of property transactions



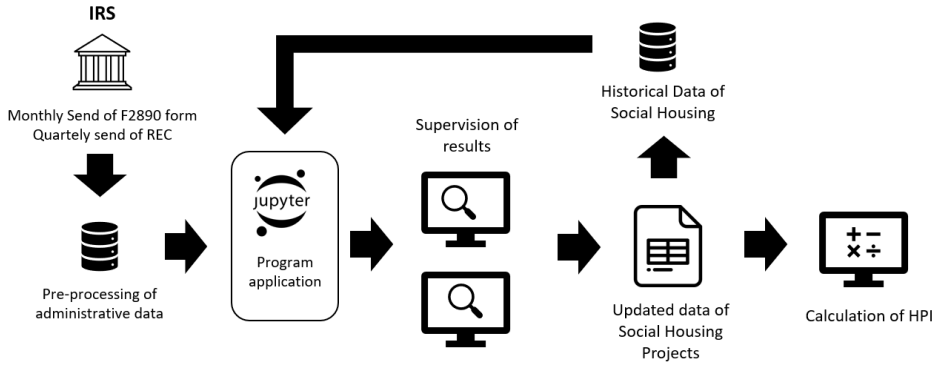
Elaborated by the authors using information from the Chilean Internal Revenue System

8 Operability in a regular basis in the Central Bank

As it was mentioned, the Household Price Index is a trimestral publication of the Central Bank of Chile. We designed a flow for the standardization of the model as a regular process during the calculation of the Household Price Index.

After the modelling process, the results of the model are supervised in parallel by two analyst who review each Block selected as social project. This validation includes a revision of news and social media to check whether the subsidized block was already inaugurated by government authorities. After that revision, results from both analysts are compared and then aggregated to a database of social blocks. Then, subsidized projects identified are excluded for the calculation of the HPI. This process occurs each trimester.

Figure 2: Operational Process of Social Project Detection



Elaborated by the authors

9 Conclusion

In this paper, we proposed a supervised classification approach using Machine Learning (ML) and statistical algorithms to automate the detection of social housing transactions in the Household Price Index dataset used by the Central Bank of Chile. We constructed new variables from the administrative records and compared the performance of three different ML models of binary classification: regularized weighted lo-

gistic regression, kernel weighted support vector machines, and regularized neural network using dropout regularization methods.

Our results showed that the Random Forest model outperformed the other models in terms of accuracy, AUC, and recall scores. We demonstrated the operability of this approach as an usual process of the Central Bank of Chile, which highlights the potential benefits of applied ML for Central Banks. Overall, our research shows that ML techniques can be effectively applied to the challenges brought using big datasets in Central Banking, providing practical solutions to data management issues, and improving the accuracy and efficiency of important economic indicators such as the Household Price Index.

Further research could be explored to improve and extend the results of this study. For example, one area of further investigation could be to explore the use of other ML to improve the accuracy and efficiency of the classification approach. Additionally, the proposed approach could be expanded to include other types of social projects or anomalies in the data. Finally, it could be interesting to explore the use of Natural Language Processing (NLP) techniques to analyze unstructured data sources, such as news articles or social media, to improve the forecasting accuracy of the Household Price Index. Overall, our research provides a starting point for further investigations into the use of ML techniques in Central Banking, highlighting the potential benefits of these tools for data management and decision-making processes.

References

D. Araujo, B. Giuseppe, J. Marcucci, R. Schmidt, and B. Tissot. Machine learning applications in central banking. *IFC Bulletin*, 57, 2022.

J. Balsa and J. Vásquez. Índice de precios de vivienda banco central de chile 2022. *Estudios económico estadístico*, 139, 2023.

I. Basheer and M. Hajmer. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods*, 43:3–31, 2000.

N. Benoudjit and M. Verleysen. On the kernel widths in radial-basis function network. *Neural Processing Letters*, 18:139–154, 2003.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Central Bank of Chile. Household finance survey. Technical report, Central Bank of Chile, Santiago, 2022.

Alexandru Coser, Monica Maer-Matei, and Crisan Albu. Predictive models for loan default risk assessment. *Economic Computation and Economic Cybernetics Studies and Research*, 53:149–165, 06 2019. doi: 10.24818/18423264/53.2.19.09.

N. Cristianini and B. Scholkopg. Support vector machines and kernel methods: the new generation of learning machines. *AI Magazine*, 23(3), 2002.

A. Cutler, D.R. Cutler, and J.R. Stevens. Random forests. In C. Zhang and Y.Q. Ma (eds.), *Ensemble Machine Learning*, pp. 157–175. Springer, New York, 2012.

S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35:352–359, 2002.

H. Faris, I. Aljarah, and S. Mirjalili. Training feedforward neural networks using multi-verse optimizer for binary classification problems. *Applied Inteligence*, 45(2):322–332, 2016.

R. Flores and J. Pérez. The housing price index for chile: Methodology and results, 2015.

P. Hidalgo. *La vivienda social en Chile y la construcción del espacio urbano en el Santiago del siglo XX*. Ril editores, 2019.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, 2001.

V. La Serra and E. Svezia. Statistical matching for anomaly detection in insurance assets granular reporting. Working Papers 22, IFC, 2022.

P. Maddaloni, D. Continanza, A. Del Monaco, D. Figoli, M. di Lucido, F. Quarta, and G. Turturiello. Stacking machine-learning models for anomaly detection: comparing anacredit to other banking datasets. *Questioni di Economia e Finanza*, (689), 2022.

S. Maldonado, R. Weber, and J. Basak. Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences*, 181:115–128, 2011.

F. Pedragosa, G. Varoquax, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, pp. 2825–2830, 2011.

A. Salvi Del Pero, W. Adema, V. Ferraro, and V. Frey. Policies to promote access to good-quality affordable housing in oecd countries. Social, Employment and Migration Working Papers 2, OECD, Paris, 2016.

N. Srivastava, G. Hinton, I. Krizhevsky, and Salakhutdinov. Dropout: A simple way to prevent neural network from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

Y. Wang, Y. Zhang, Y. Lu, and X. Yu. A comparative assessment of credit risk model based on machine learning - a case study of bank loan data. *Procedia Computer Science*, 174:141–149, 2020.

L. Ying. Research on bank credit default prediction based on data mining algorithm. *International Journal of Social Sciences and Humanities Invention*, 5(6):4820–4823, 2018. doi: 10.18535/ijsshi/v5i6.09.