Designing Policy Recommendations to Reduce Home Abandonment in Mexico*

Klaus Ackermann
Monash University
klaus.ackermann@monash.edu

Eduardo Blancas Reyes
Center for Data Science and Public Policy
University of Chicago
edublancas@uchicago.edu

Sue He
University of Virginia
syh2s@virginia.edu

Thomas Anderson Keller
UC San Diego
andy@keller.org

Paul van der Boor
University of Chicago
pvboor@gmail.com

Romana Khan
Northwestern University
romana.khan@kellogg.northwestern.edu

ABSTRACT

Infonavit[1], the largest provider of mortgages in Mexico, assists working families to obtain low-interest rate housing solutions. An increasingly prevalent problem is home abandonment: when a homeowner decides to leave their property and forego their investment. A major causal factor of this outcome is a mismatch between the homeowner's needs, in terms of access to services and employment, and the location characteristics of the home.

This paper describes our collaboration with Infonavit to reduce home abandonment at two levels: develop policy recommendations for targeted improvements in location characteristics, and develop a decision-support tool to assist the homeowner in the home location decision. Using 20 years of mortgage history data combined with surveys, census, and location information, we develop a model to predict the probability of home abandonment based on both individual and location characteristics. The model is used to develop a tool that provides Infonavit the ability to give advice to Mexican workers when they apply for a loan, evaluate and improve the locations of new housing developments, and provide data-driven recommendations to the federal government to influence local development initiatives and infrastructure investments. The result is improving economic outcomes for the citizens of Mexico by pre-emptively identifying at-risk home mortgages, thereby allowing them to be altered or remedied before they result in abandonment.

*This work was done at the 2015 Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago.

[1]Instituto del Fondo Nacional de la Vivienda para los Trabajadores (National Institute of Housing Fund for Workers)

KDD '16, August 13-17, 2016, San Francisco, CA, USA

 ISBN 978-1-4503-4232-2/16/08…$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939702

Figure 1: Abandoned and vandalized house (Zumpango, Mexico)

1. INTRODUCTION

Home abandonment refers to a homeowner's decision to leave their property and forego the financial investment they have already made in the home. This decision imposes negative financial and economic consequences for both individuals and mortgage-financing institutions in Mexico, as well as imposing significant negative externalities on the neighborhood. To date, home abandonment has affected more than 200,000 homes, representing approximately 5% of the homes financed by Infonavit. As the largest provider of mortgages in Mexico, Infonavit serves approximately 70% of the mortgage market in the country. It was created as a public institution for the purpose of assisting lower-income families, who cannot obtain financing from private financing institutions, with home acquisition and other housing solutions.

In Mexico, every employer registered in the National Social Security Institute is required to contribute 5% of each worker's salary to Infonavit, which administers the contributions as a savings account. When employees meet certain

Figure 2: Neighborhoods with high abandonment rate have also high crime rates, which forces the local population to gate entire streets (Zumpango, Mexico)

criteria, they are eligible to apply for a credit, typically to buy a new house.[4]

Most workers who acquire a house through an Infonavit loan come from low-income families. Faced with economic constraints, they tend to buy cheaper houses which are located at a distance from city centers, often in new housing developments in remote suburbs with limited infrastructure and limited access to services and amenities. At the time of purchase, workers may be poorly informed about the day to day realities of their choice. For example, they may fail to properly account for the time and cost involved in commuting to their employment or schooling. It is often the case that families only realize the location does not satisfy their needs after they have moved into their new home, leading to a decision to abandon the house. Previous research indicates that workers abandon their home for several reasons, including long distances to their workplaces and schools, a lack of services, finances, and security concerns.

Infonavit wants to reduce home abandonment, but the current process is reactive and often too late. Loan holders who stop making payments enter a portfolio of low performing loans. Then, payment collectors visit each home and through a lengthy process of extensions and repeated home visits the home is marked as abandoned. This process takes approximately 12-18 months. At this stage the abandoned houses may have already motivated neighbors to leave their house behind as well and neighborhood regeneration programs may be too late to be effective.

1.1 Project objective

This project has the primary objective of reducing home abandonment in Mexico. We address this in two ways:

1. Enable Infonavit to predict the risk of home abandonment for a given individual and home, and use that prediction to provide purchase advice to the individual.
2. Provide policy recommendations to the government using the structure and features of the predictive model.

The first objective is achieved by building a machine learning model that, given an individual applying for a loan and a home location, can predict the risk of abandonment. By evaluating the risk scores over the homes in the individual's choice set, Infonavit can provide advice by identifying which homes would have a lower risk of abandonment in the future. The second objective is to understand the impact of specific location features on home abandonment. Extensive research carried out by Infonavit indicated that distance to employment, and access to facilities such as schools, hospitals and grocery stores were instrumental in the home abandonment decision. What was less well understood was the relative impact of each of these factors, and the maximum distance thresholds that workers were willing to travel. Access to this information will enable Infovanit to improve the location of new housing developments, give data-backed recommendations to the federal government to influence public policy, and provide guidance to improve the home location decisions of Mexican workers.

We believe that these two actions will help prevent the establishment of settlements with high abandonment rates as portrayed in figure 1 and 2. Currently, revitalization eforts in areas with high abandonment rates only begin after the problem has become apparent for years.

2. PROBLEM FORMULATION

Overall, we formulate the abandonment risk prediction problem as a binary classification problem where the outcome variable is whether a person abandons their house. We model this outcome for each year after the loan is granted, either until the end of the observation period or until the home is abandoned.

An alternative is to model the percentage of abandonment for a specific colonia[2]. This would not serve our objectives since the analysis would no longer be focused on the individual decision of a person to abandon their house. Estimation at the colonia level would also require aggregating individual level data to a representative average person, and we would lose the information from individual level variation. Furthermore, as the specific addresses of houses are unknown for a large portion of the data, attempts to geocode these houses were unsuccessful due to the low quality of the address fields[3].

3. DATA SOURCES

To investigate the factors that contribute to home abandonment, we combined data from multiple sources. The primary level of observation is at the individual home owner. This is supplemented with multiple data sources that capture the location characteristics of the house.

3.1 Loans data

The primary data provided by Infonavit includes personal, loan, and house characteristics for every loan granted in the last 20 years. The personal characteristics include demographics such as age and marital status, and financial infor-

[2]Neighborhood

[3]It's important to note that while we are predicting at a home level (to asess which homes are likely to be abandoned in the next year) many of our features are at different levels of aggregation, from the home level up to the municipality level

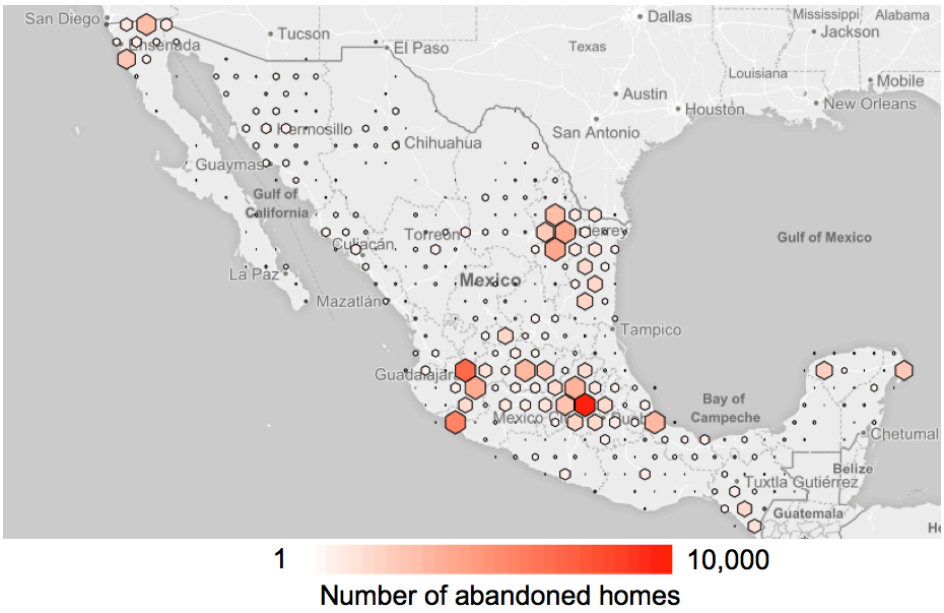

Figure 3: Geographic distribution of home abandonment in Mexico

mation such as wage and the value of accumulated savings. The data also includes a *Risk Index* computed by Infonavit which estimates the risk of employment loss for every credit-holder. The information on the terms of the loan included the interest rate, the value of the loan, and the value of any subsidy received.

We transformed the data so that the unit of observation is at the house level rather than the individual loans, since it is possible for a family to have two loans for a house. We combined their savings and assets so that each observation is representative of one home with its associated characteristics. We excluded joint loans that had mismatched years or prices for the same house. Loans cofinanced with private banks, loans that were granted for the purpose of home improvement, and loans with entries that did not match up with other data sources provided by Infonavit (such as guidelines corresponding to the age of the loan holder) were also excluded. After this cleaning process, we have 4.1 million observations at the house level.

One field in the data that is often missing is the exact location of the house. Unfortunately, geo-coordinates were only available for loans granted in the last 7 years (since 2008), and attempts to geocode the rest of the houses were unsuccessful due to the low quality of the address field. To deal with this practical constraint, we matched loans to colonias[4] using different approaches. A detailed description of this process is described in section 4, Location Matching.

3.2 Approximation of the labeled data

Home abandonment is recorded as a flag in our data but the exact date of abandonment is often unknown. Our loan data (where the abandonment label is located) presents a snapshot of the state of abandonment as of May 2015 and the abandonment flag tells us whether the home is abandoned as of that date.

[4]neighborhoods in Mexico

The abandonment flag can be triggered by different circumstances. The loan repayment occurs through the social security system directly from the employer. Payments cease when a person loses their job, and continue with the uptake of new employment. It is also possible that the person never regains any form of formal employment and joins the informal sector in Mexico. In the first case, a house will be marked as abandoned or vandalized when an inspector visits a colonia. In the second case of no payments, after a grace period an inspector visits the house and tries to locate the owners. Both circumstances are equally devastating for the livelihood of their former owners as well as the remaining occupants in the colonia.

Approximate dates of home abandonment were calculated based on supplemental payment data provided by Infonavit, which recorded payments made for every active loan during the 2010-2014 time period. We assumed that the first interruption in the payments represented the date of home abandonment. There were 4 types of events which were used to identify an interruption in payments:

1. Date of the application for loan extension granted (59%)
2. Date of the first interruption in the payment stream with continued payments after (11%)
3. Date of last payment received (11%)
4. Date the home is marked as abandoned but no single payment entry, date set to 2009 (5%)

For 14% of the data we could not find a pattern and excluded these observations from the study.

3.3 Housing survey data

Infonavit also shared the results of home surveys (ECUVE[5]) conducted by licensed inspectors. This measured housing quality at the municipality[6] level. ECUVE provides an in-

[5]Evaluación cualitativa de la vivienda y su entorno (Quantitative evaluation of housing and its environment)

[6]Municipality is the administrative division of states in Mexico

dex of housing quality based on factors such as construction quality, reliable water and power supply, local access to schools and hospitals, and availability of parks and markets.

3.4 Business, school and hospital location data

Previous research conducted by Infonavit indicated that distance to employment was among the top factors that drive home abandonment. To account for this, we incorporated data from DENUE[7] in our analysis. Every five years, INEGI[8] conducts a census of businesses in Mexico, covering businesses from small grocery stores to multinational corporations. For each business, information such as industry, number of employees, owner, address, and geolocation coordinates is recorded. The data also includes schools and hospitals. The study is updated annually.[9]

3.5 Municipality data

Finally, we included extensive data on the folliwing municipality characteristics:

- Number of Homicides
- Natality
- Mortality
- Natural disaster incidences
- Years of schooling (from the Population Census)
- Households statistics (from the Population Census)
- Literacy rate (from the Population Census)
- Healthcare coverage (from the Population Census)
- Number of vehicles and passenger buses

All the municipality level data comes from INEGI (except for the natural disasters incidence data, which comes from Desinventar [1]). INEGI's datasets are publicly available for download [2].

3.6 Data limitations and project scope

Most of the data available to us is collected annually, except for the Population Census, which occurs every five years. This represented a challenge for feature creation, described in the next section. The loans dataset spans the last 20 years, but house coordinates were only available for loans granted after 2008. Given that location features were critical to our objectives, the scope of the project was limited to loans granted from 2008 to 2015. A timeline of the data is presented in Figure 4.

4. LOCATION MATCHING

One of the biggest challenges for our project was to geolocate granted loans. Since our model depends on spatial-temporal features (e.g. number of schools within 5 Km), we need to obtain coordinates for each loan to be able to compute those features.

Our loans dataset contains more than 4.1 million rows, but location information (coordinates and/or address) is available for only 2 Million (49%) observations. To geolocate address data, a common approach is to obtain geographical coordinates using a geocoding service. In the U.S., for example, the Census Bureau has a geocoding API[7] that given an address returns its coordinates.

Unfortunately, there is no official service for geocoding addresses in Mexico. One of the alternatives is to use Google Maps Geocoding API[6]. The first limitation with Google's service is the rate limit of 2,500 requests per day. Furthermore, the service is not as accurate as in the U.S., mostly because many Infonavit housing developments were built recently and are still not mapped by Google.

With that in mind, we decided to geolocate at the colonia-level only and group together all houses in the same colonia in a point located at the colonia's centroid.

4.1 Matching loans to colonias

The first step to match loans (4.1 million) with their corresponding colonia, is to filter those that have at least address information (some have coordinates).

The next step was to match loans that had geographical coordinates with their corresponding colonia, this was done using PostGIS spatial queries. We matched 740 000 loans using this method. After that we tried locating non-matched loans by finding their closest matching-loan and assigning the same colonia, this process was limited to a distance up to 1km.

The next step to attempt to locate more loans was to use the address field. For each non-matched loan, we compared it to the matched ones and assign the same colonia if the address was the same. We made this comparison using a simple string comparison. Due to time constraints we didn't implement a more robust approach such as entity resolution. In the end we were able to geolocate 2.4 million records (58.5% of total loans).

5. FEATURES

Once we received and cleaned the data (preprocessing such as translation and deflation to account for time variation) and finished the location matching process, we proceeded to build features. Each type of features required a different process, which is described in the following subsections.

5.1 Municipality features

Except for the Population Census, all municipality features are collected yearly. Since some years were missing, the missing values were imputed using linear extrapolation (e.g. to impute a feature for 2010, extrapolate using data from 2008 and 2009). Another approach could be to use interpolation (e.g. to impute a feature for 2010, interpolate using data from 2009 and 2011), this method was rejected since using data from the *future* (in 2010, we don't know the value for 2011 features) would violate our testing and training split paradigm.

Municipalities in Mexico do not have a uniform distribution in terms of area or population. For example, the most populated municipality (Ecatepec) has more than 1.5 million inhabitants, whereas the least populated one (Santa Magdalena Jicotlán) has only 102 inhabitants, in term of land area, the biggest municipality (Ensenada) has an area of roughly 52,000 km^2 and the smallest municipality (San Lorenzo Axocomanitla) has only 4.2 km^2. To account for this, features involving discrete counts were normalized over

[7]Directorio Estadístico Nacional de Unidades Económicas (National Statistical Index of Economic Units)

[8]Instituto Nacional de Estadística y Geografía (National Institute of Statistics and Geography)

[9]Every five years the study is carried out from scratch and every year only a subset of the data (80% of the GDP) is updated. For detailed methodology information refer to [3]

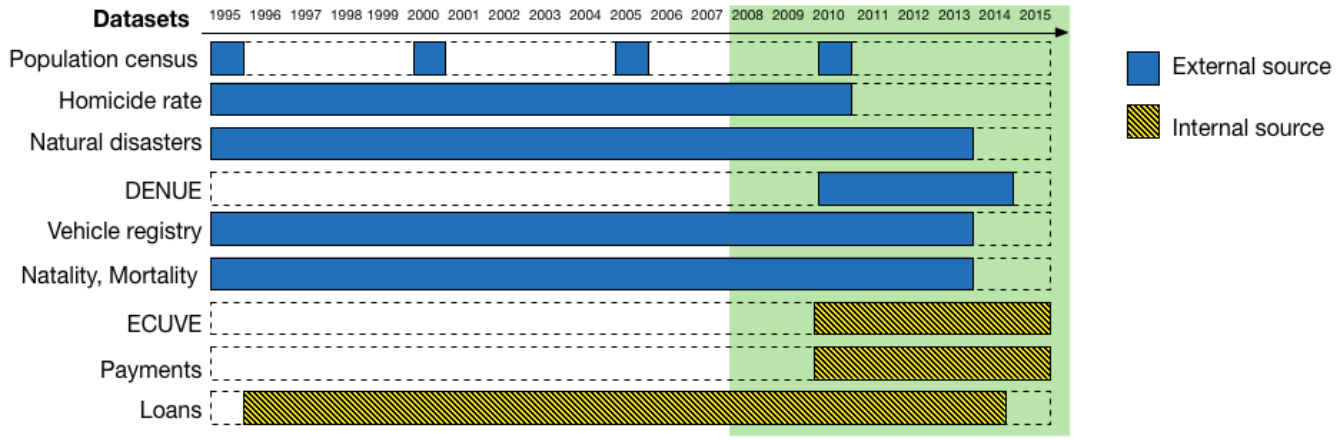

Figure 4: Data timeline summarizing the data sources used and project scoping (2008-2015)

100,000 inhabitants (e.g. number of homicides, births, deaths, households).[10]

5.2 Location features

Location features were critical to our project. They were of strategic value to provide actionable policy recommendations to our partner. Our personal impressions supported this during a site visit to a neighborhood with high abandonment (Zumpango, Mexico). The requirement is to have features that capture the state of the local services and infrastructure in a comparable way across Mexico. We chose to aggregate several types of businesses separately for each colonia, and at various radii (from 0 to 5 km, 5 to 10 km, up to 45 to 50 km). The number of employees for businesses within each concentric circle was also counted. It is an approximation of the actual infrastructure, nevertheless it should be representative across the state.

The accurate distance calculation between approximately 50,000 colonias, spanning some 4 million businesses across Mexico is an extensive endeavor. As we had no road network data to calculate the actual walking or driving time we relied on the air distance corrected by the spheroid distance. Without accounting for the curvature of the earth we might introduce noise in our geometrical aggregation bins, when a business might be counted differently depending on the distance. The challenge that arose was that a single feature calculation for a colonia took between 1 to 5 seconds, depending on the density of businesses. We developed a plsql extension parallelsql[11] for PostgreSQL that allows these distance calculations to be executed in parallel, decreasing the required calculation time by the number of cores available to the database. The input data is partitioned by the primary key in split up in chunks of colonia groups. A load balance implementation submits the chunks to the cores of the database and waits until a sub part is finished and submits the next one until all features are calculated.

5.3 Personal and loan features

Personal and loan features were used for each observation with new aggregate features created to characterize each colonia, taking the colonia average, minimum and maximum for both personal and loan features.

5.4 Merging features

Following the feature engineering step, a master feature table was built to contain all features, with each row having one single loan-year and the following features:

- Credit-holder characteristics
- Loan characteristics
- Average credit-holder characteristics (colonia level)
- Average personal characteristics (colonia level)
- Municipality features
- Location features

6. MODELING

6.1 Model objectives

Our model was created to answer the following question: What is the risk of abandonment for an existing loan in the next year?

We trained a variety of machine learning models on different years and then tested on active loans in the following year (e.g. training from 2008 to 2014 and testing on loans active during 2015). The algorithms tested were Support Vector Machines, Random Forests, AdaBoost, and Logistic Regression.

6.2 Accounting for loan duration

In order to accurately model the effect of loan duration on home abandonment and to be able to easily split our training and testing sets correctly, we duplicated records for each loan from the granting time until the abandonment year (e.g. for a loan granted in 2008 but abandoned in 2010, there will be 3 rows, one for each year), with the most recent row having the abandoned flag as true.

6.3 Generating our training and test sets

To evaluate our models, we had to ensure that we were not violating time-dependent knowledge restrictions between our training and test set. This means using the known economic conditions at the time of *evaluation* rather than at *testing*

[10]INEGI 2010 Population census
[11]https://github.com/k1aus/parallelsql

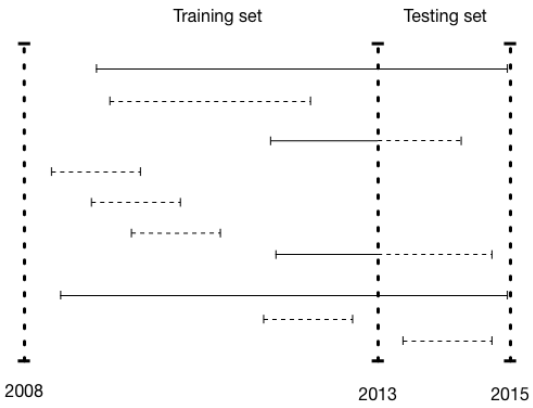

Figure 5: Training (2008-2013)/Testing (2014-2015) splitting example

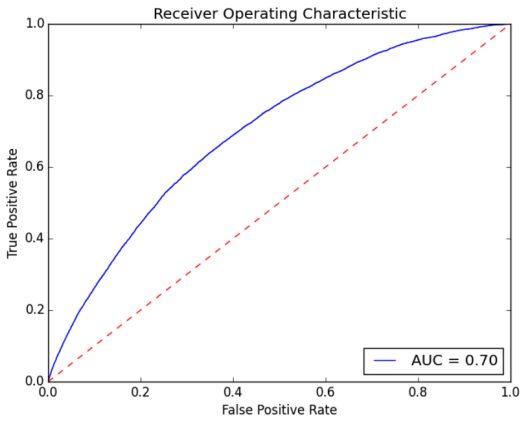

Figure 6: ROC curve for our best model

Risk Index
Years since loan granted
Loan sales price
Daily wage
Loan account value
Colonia minimum sales price
Loan interest rate
Loan holder age

Table 1: Top personal/loan predictors

time. For example, if we wanted to predict home abandonment for 2015, the only information available would be up until 2014, since later information would be using data from the future.

Figure 5 shows a data split using data from 2008 to 2013 for training and 2014 to 2015 for testing. A continuous line means the observation is considered to be inhabited, while a dashed line means the observation is considered abandoned.

To train our models we used scikit-learn[5], a Python library that contains many algorithms for binary classification.

7. RESULTS

Because of the time restrictions on accurate home location data, and our approximation of dates of home abandonment (explained in previous sections), our final models was trained using data from 2008 to 2014 and tested on 2015. This resulted in a 95 to 5 negative to positive class ratio. To tackle this imbalance we oversampled the positive class in our training set to have 50% positive observations, while leaving the test set untouched. Our main evaluation metric was chosen to be AUC given it's interpretation as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, which could be directly applied to the ranking of risk for multiple loans.

7.1 Model performance

Our best performing model was a Random Forest which achieved an AUC of 0.70 as shown in Figure 6. With a 0.5 threshold, the model captured 55% of abandoned houses. Given the inherent imbalanced nature of the data, this model produced 266,670 false positives in comparison to 4602 true abandoned houses.

While the predictive power of our model is limited, the top predictors (see subsection 7.2) provide substantial deep insights into the data and provide guidance on where resources need to be allocated to address the problem of home abandonment (see section 8.2).

Several factors contribute to the model's limited performance (see subsection 7.3), and suspect that the vast majority of them relate to the fidelity and quality of the data. To address this, we provided Infonavit with a series of data collection recommendations (see section 8.1) to aid in the improvement of the performance of the current model.

7.2 Top predictors

Trained models consistently show the same sets of features as important to predict home abandonment. Table 1 shows the top features regarding inherent personal and loan characteristics and Table 2 shows the top features that can potentially be impacted through policy.

7.3 Model limitations due to data quality

Determining accurate coordinates of homes was a major limitation we encountered while attempting to precisely compute location features (such as distance to nearest employment areas). Since we only had exact coordinates for about 740,000 of 4.1 million loans, the most accurate location we could calculate for homes with missing location data was the center of the colonia in which they were located. Using various fuzzy matching methods we retrieved colonia locations for 2.4 million loans, still missing around 1.7

ECUVE index
Number of restaurant employees within 5 km
Number of employees within 5 km
Number of businesses within 5 km
Number of hospitals within 5 km
Loan-holder age * Number of schools within 5 km
Number of churches within 5 km
Municipality Passenger buses over 1000 inhabitants

Table 2: Top predictors that can be improved through policy changes

million. This location approximation introduced a substantial amount of noise into our location features since colonias can stretch multiple kilometers and we had no indication of where a given home was located inside each colonia.

The poor location information created two problems for our analysis. First, we were unable to compute actual distances to nearby businesses in meters, and instead opted to count the number of businesses within distance rings. Second, since we were unable to locate an individual loan within a colonia, we could not compute intra-colonia features which could have helped our model distinguish between homes (especially those with similar loan information) in a given colonia.

Related to the distance features, our model strongly depends on DENUE. This process to update this dataset is carried out every 5 years, so it is possible that the current model does not capture all the economic changes that occur year to year.

Limitations related to data acquisition include our inability to obtain public transportation information (e.g. location of bus stops) and an inability to obtain accurate employment information and location for each loan holder.

Another limitation is the uncertainty of abandonment date within official Infonavit records. Home abandonment is a complex process, so establishing the exact timestamp of when a home was abandoned would is difficult if not impossible without all houses being checked consistently. The only timestamp we had originally from the raw data was the year in which the loan was granted. For our models, the abandonment year needed to be estimated from the loan payment history data.

7.4 Prototype

The predictive model (Figure 7) was deployed as a web application which estimates the risk of abandonment for the next year based only on personal characteristics and a selected colonia for the home. The process to make a prediction is as follows:

1. User selects a colonia
2. User inputs personal characteristics (age, *Risk Index* and daily wage)
3. The web application retrieves the loan characteristics as the average for that colonia
4. Municipality and location features are retrieved from the database
5. The aggregated features are run through the trained model, and the application displays the prediction to the user, along with a map and a summary of the factors driving home abandonment

8. CONCLUSIONS

Our model made a modest but substantive improvement in home abandonment prediction compared to a baseline model of Infonavit's internal *Risk Index* alone. Before our project, the institute had no way to estimate the risk of home abandonment at the loan level. Taking advantage of the insights found, Infonavit will be able to take the following actions:

1. **Advise** Mexican workers when they are applying for a loan, so that home abandonment can be prevented at the loan origination time. Using our model, Infonavit can estimate the risk of home abandonment given personal characteristics and location. With this information, they will be able to suggest alternative locations better suited to the loan applicant and decrease the risk of abandonment. It is important to mention that Infonavit cannot deny loans to formally employed workers by law, so our model results will only be used to provide guidance to potential loan holders and to help them make a better location decision.
2. **Prevent** the spread of home abandonment in certain neighborhoods. By identifying locations which are at risk of abandonment, pre-emptive action can be taken to reduce such risk (e.g. intervention programs).
3. **Improve** planning for new developments. Infonavit does not build houses, but purchases them from development companies. Taking advantage of our work, Infonavit will have enough information to asses the risk associated with a new development in a certain location.
4. **Influence** public policy at the federal level. Solving the problem of home abandonment in Mexico involves many stakeholders such as Infonavit and the federal government. Based on the results of our work, Infonavit will be in a better position to influence public policy and to prioritize changes based on our findings.

8.1 Data collection recommendations

To improve the model's ability to accurately predict abandonment, we recommend Infonavit make certain improvements in the data collection process. Most important is to collect precise geographic coordinates for all houses. Since location related features are important predictors of home abandonment, it is critical to improve the quality of this data.

The second recommendation is to start collecting the loan holder's employment location and school location data. According to previous Infonavit research, commuting time is a top reason for home abandonment, so incorporating this information in the model will provide new insight and may have a significant effect in the model performance.

The interval ECUVE score created at Infonavit is a top predictor of home abandonment but there were many missing values in some component variables (ECUVE measures different variables such as transportation, access to water, parks, markets and aggregates them to provide an overall score). Some were unsuitable for use in our models. Given the promise of the overall ECUVE score as a predictor, it is important to improve the quality of the ECUVE data. One priority could be to increase the sample size to include more municipalities in the country.

Our last recommendation is in regards to the time records. Since the desired output of a model is the risk of abandonment with respect to a change in time (in the current model we predicted abandonment in the next year), it is important to improve the temporal data quality, meaning that Infonavit needs to record important events with more precision such as month of loan granting and if possible, the month of abandonment. At the start of the project the only raw timestamp recorded at Infonavit was the loan granting year. We thus needed to approximate the date of abandonment in order to accurately consider the change in our features over time (e.g. number of jobs within 10 km in 2012 versus 2013).

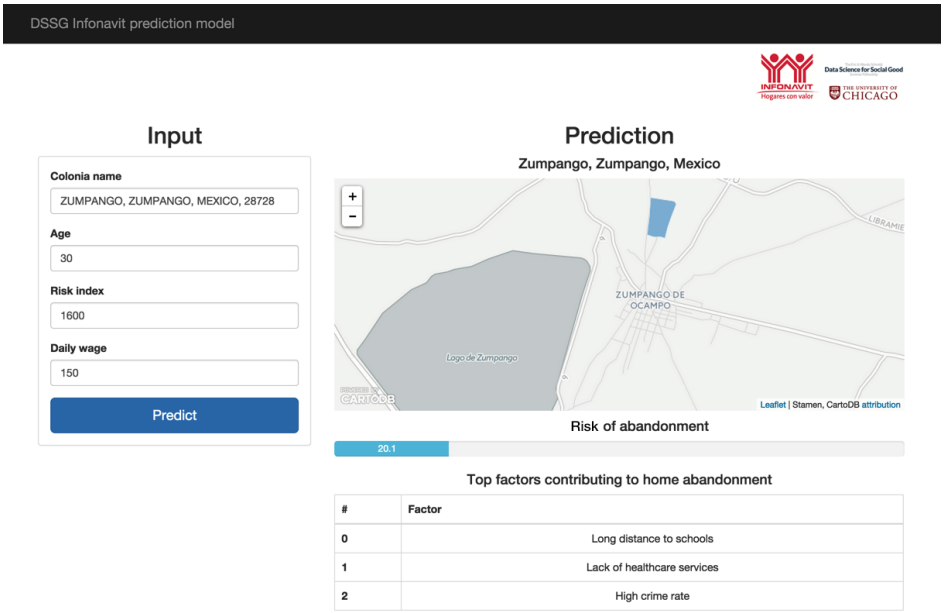

Figure 7: Prototype screenshot

This introduced noise in our model, thus severely affecting the performance.

8.2 Public policy recommendations

Infonavit cannot change public policy as an individual actor, but has a close relationship with the federal government and can advise the them and influence other players to do so. With a considerable spectrum of possible policy changes yet limited resources to deploy them, having a data-backed starting point is crucial to tackling home abandonment.

Our recommendation pinpoints the need for a focus on local rather than regional growth. Our model found that the critical distance for services is within 5 km, so creating large centralized employment and educational centers at distances further than 5km from colonias would not have a significant impact on reducing home abandonment.

Home abandonment is not an intermittent phenomenon across time, instead it spreads inside neighborhoods, so we encourage Infonavit to allocate resources for urban regeneration and intervention programs in neighborhoods that already have higher percentages of abandonment.

9. ACKNOWLEDGMENTS

This project would not have been possible without the financial support provided by the Schmidt Family Foundation, making not only this project possible, but also another eleven in the Data Science For Social Good (DSSG) program.

We would also like to thank DSSG staff for making the program possible and for all the support provided throughout the summer. Their hard work inspired us every day to do our best and to make an impact.

Finally, we are grateful to Infonavit for trusting and actively working with the team to provide valuable insights. We hope our results help the institution better accomplish its social mission and improve the quality of life for many families in Mexico.

10. ADDITIONAL AUTHORS

Rayid Ghani (University of Chicago, email: `rayid@uchicago.edu`) and José Carlos González (Infonavit, email: `josecarlosgonz@gmail.com`).

11. REFERENCES

[1] Desinventar. Inventory system of the effects of disasters. http://www.desinventar.org/en, 2013. [Online; accessed 09-September-2015].

[2] INEGI. Descarga masiva. http://www3.inegi.org.mx/sistemas/descarga/, 2013. [Online; accessed 09-September-2015].

[3] INEGI. Directorio estadístico nactional de unidades económicas. http://www.inegi.org.mx/est/contenidos/proyectos/denue/presentacion.aspx, 2015. [Online; accessed 09-September-2015].

[4] Infonavit. Qué hago para cumplir. http://portal.infonavit.org.mx/wps/wcm/connect/infonavit/patrones/mis_compromisos/que+hago+para+cumplir. [Online; accessed 09-September-2015].

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Google. The Google Maps Geocoding API. https://developers.google.com/maps/documentation/geocoding/intro, 2016. [Online; accessed 15-January-2016].

[7] United States Census Bureau. Geocoder. http://geocoding.geo.census.gov, 2010. [Online; accessed 15-January-2016].