MDPI

Article

Predicting Lending in Arizona, Ohio, and Kansas Using Local Cimatological Data

Naomi Liftman[1], Alison Gu[1], Michele Sezgin[1], Jane Andrews[1], Kate Phan[1]

[1] Smith College 1 Chapin Way, Northampton, United States; bbaumer@smith.edu
* Correspondence:
† Current address: Updated affiliation
‡ These authors contributed equally to this work.

Version May 9, 2023 submitted to Water

1. Project Overview

While weather affects the general population's day to day life, there may also be effects of weather on corporations' and enterprises' interactions with that general public. We studied bank data on agricultural, real estate, and commercial real estate loaning along with weather data for Ohio, Arizona, and Kansas to prove the capabilities of this type of study. While our findings were suboptimal at best, with little evidence that weather affects loan habits for banks in these areas, our study does prove that this is a viable research opportunity when expanded to a larger team, more time, and the entire country.

2. Introduction

Since the invention of the steam engine and its widespread use starting in the late eighteenth century with the advent of the industrial revolution, humans have made changes to ecosystems, watersheds, and broadly the environment on a scale previously unimaginable [1]. The Cleveland Federal Reserve Bank has recently focused attention on the effects of climate change on their work. The goal of the bank, at large, is to research, monitor, strengthen, and stabilize the economic performance and monetary system of the United States, with a focus on Ohio, western Pennsylvania, eastern Kentucky, and the northern panhandle of West Virginia, which they oversee.

Our team is working with CFRB, specifically Thealexa Becker and Catherine Chen, to assist them in their goals of analyzing global warming and its effects on loan types. CFRB had previously been working on this project, but resources were shifted towards other efforts. Our team was given the task of addressing this gap by working on a proof-of-concept version of this project, wrangling the data and testing out methods of analysis for when CFRB can pick up the project again.

There is a lack of analysis on the past and probable future relationship between climate and loan data in the United States over the past two decades, for several reasons. Carbon-reliant corporate interests have protected their profits by obstructing progress in climate change recognition and solutions [2]. Further, there are currently no industry standard models or practices for integrating climate risk into financial risk management, and no benchmarking datasets for testing models. Most climate/weather data is extremely large, so providing concise estimates of relevant climatological factors can be difficult. Lastly, most of the analysis in this field is done by third party consulting groups, who do not make their findings publicly available.

More broadly, this lack of research is concerning due to the risks banks are exposed to as climate change worsens. Many industries are already facing more difficulties providing consistent and reliable productivity as climate change progresses, which will in turn negatively affect those financing them and the stability of the financial system as a whole. Agriculture is one notable example; crop losses

due to climate events have been increasing over recent decades [3], a concerning trend. In recent years, banks have become more vocal about the risks climate change exposes them to, with one of the main concerns being that physical impacts of climate change will "lead to substantial losses to investors and could spread to other sectors of the economy and undermine the stability of the financial system" [4].

Our goal for this project is to analyze climate and weather data in relation to loans made by the Federal Reserve, which aligns with their mission. The Federal Reserve is invested in maintaining the stability of the U.S. economy, an interest that anthropogenic climate change puts at risk. Better understanding and potentially taking preventative measures to avoid adverse economic impacts of the climate crisis therefore is highly related to the goals of CFRB.

3. Data Description

There are two main data sources that we will be working with in this project, the Consolidated Report of Condition and Income, or Call Report data from the Federal Financial Institutions Examination Council [5], and the Local Climatological data from the National Oceanic and Atmospheric Administration [6]. The Call Report data contains the addresses of approximately 6500 insured commercial banks in the U.S, each of which has a unique identifier. The NOAA data contains geographic coordinates of 2464 weather stations in the U.S. The portion of the larger datasets we used for this project includes that of banks and weather stations in three states selected by our sponsor: Ohio, Arizona, and Kansas. The Call Report includes details regarding financial transactions a bank completed for each quarter of a year. These include the types of loan, amounts of loan, and other relevant information. The NOAA data include detailed climate information collected from different weather stations in the fourth and twelfth districts, such as hourly humidity, precipitation, and air pressure. Each weather station generates around 12,000 lines of climatological data every year. The NOAA climate data dates back to 1901, but we will only be examining data from the years 2000-2022 for our purposes. Both datasets are fairly high dimensional in their raw state. The NOAA climate data contains 125 variables, while the call report contains 165 variables potentially relevant to our analysis.

The Federal Deposit Insurance Corporation is responsible for making sure that financial institutions such as banks adhere to reporting requirements, as this allows for transparency and accountability. Since the 2008 financial crisis Call Reports have become a far more common manner of analyzing Banks and their loan patterns. Then our second dataset which we will be merging is from NOAA, who is a scientific and regulatory agency that works within the US Department of Commerce. They monitor oceanic and atmospheric conditions, chart the ocean and deep seas, and forecast weather conditions. They also manage fishing and protection of marine mammals and endangered species in the U.S. exclusive economic zone, and all of these goals of NOAA have culminated in the historic weather data that we will be analyzing.

Some of the slight limitations of the data is that the Call Report and NOAA dataset are missing data in many areas, which is intentional for the Call Report but not for the NOAA data, according to our sponsor. While not a limitation, an added layer of complexity is that the Call Report is difficult to download and merge. Similarly, the downloading process for the NOAA data is a little complicated as it takes at least 24 hours to download and requires about 30GB of storage on our personal computers.

4. Detailed Methodology

To understand how severe weather events affect loans of banks in Ohio, Kansas, and Arizona, we used two datasets: NOAA and Call Report. NOAA data was downloaded from the National Oceanic and Atmospheric Administration's website through their Climate Data Online system. To retrieve a list of available weather stations, we use the function isd_stations() from the rnoaa package. isd_stations() includes all geographic information of current as well as historical weather stations in the Climate Data Online system. This function also updates information on NOAA stations every day. The Call Report data was downloaded from the Federal Financial Institutions Examination Council's website through their Central Data Repository's Public Data Distribution system.

We worked with our sponsor to determine which metrics to use from each dataset. For the Call Report data, we used specific variables that they requested. Five of these variables refer to commercial real estate, three refer to real estate; and the remaining two refer to agricultural loans.

For the NOAA dataset, we looked at severe weather events, which are determined based on NOAA's own definitions. NOAA and the National Weather Service (NWS) have published a storm data preparation directive with descriptive criteria for classifying fifty-four types of severe weather events. Using these definitions, we determined that the variables available in the NOAA data would allow us to potentially detect extreme heat, heavy snow, dense fog, smoke, extreme wind chill, and high wind. From there, we wrote a script that takes in annual weather data from a station and detects how many events of each type occurred in that time period, split into quarters. These results were compared with NOAA's severe weather events database to check for validity in our detection methods. This was necessary to aggregate the NOAA data up so that it could be analyzed alongside the call report data, which uses quarter-years as its time interval.

For the NOAA data analysis, we assumed that the data was accurate and had been collected and processed using appropriate scientific methods. Limitations of the weather station equipment used are outlined by a NOAA-published user manual, which describes how equipment may have difficulty distinguishing types of precipitation, among other limitations [7]. Additionally, we assumed that the weather patterns observed during the period being analyzed are typical of the larger weather patterns in the region. There are a limited number of weather stations per state, and so we had to hold the assumption that the weather observations taken at these stations could collectively describe the weather conditions throughout the state as a whole. The existence of microclimates, which NOAA defines as "The climate of a small area such as a cave, house, city or valley that may be different from that in the general region" [8], indicate that this assumption, while necessary, may not always be correct. We also made the assumption that there are no significant outliers or anomalies that could skew the results. These assumptions help us to interpret and draw conclusions from the weather data, but it is important to be aware of uncertainties in the data.

The unit of observation of the Call Report data is bank per quarter/year and that of the NOAA data is weather station per quarter/year. To merge the two datasets together, we matched each bank with the nearest available weather station using geographic coordinates. Since the Call Report data only contains bank addresses, we first used the Texas A&M Geoservices to retrieve the longitudes and latitudes for each bank. Then, for each bank in each time period (year and quarter), we found the nearest weather station using an existing function nn2() from the RANN package. Most banks are matched with weather stations that are less than 40 miles away. Additionally, since the number of NOAA stations changes over time as some disappear and some are added later on, we made sure that our function updated the nearest weather station for each bank per year/quarter. For example, bank A is matched with station B from 2010 to 2012, but station B stops recording weather data after that. Then, from 2013 onwards, bank A is matched with station C, which is the nearest weather station. The same goes for newly added stations. In this case, bank A is matched with station C from 2013 to 2015, but in 2016, a new weather station, station D, starts recording weather data and station D happens to be closer to bank A geographically than station C. Bank A is then rematched with station D, although station C is still available.

4.1. Modeling

To understand how loaning is changing over time, we constructed our dependent variable two ways:

1. The percent increase/decrease from the previous year,

   Percent Change $\text{Loans}_t = \frac{\text{Loans}_t - \text{Loans}_{t-1}}{\text{Loans}_{t-1}}, t = 2, \ldots, n$ (Note: this is coded as a 100% increase if the previous year was 0 loans and the next year was non-zero and vice versa).

2. Classes corresponding to increase (1), decrease (-1) or no change (0).

We regressed these these two dependent variable constructions for real estate, commercial real estate, and agriculture against current and lagged hourly mean and standard deviation for wind speed, temperature, and precipitation. Additional variables included extreme heat, dense fog, smoke, extreme wind chill, and high wind (heavy snow events were also measured but no stations had any heavy snow days). Values were lagged for one, two, three, and four quarters.

4.2. Elastic Net Regression

We used only the constinuous (percent change) outcome for this model. Our alpha values for elastic net regression were 0 (ridge regression), .25, .5, and 1 (lasso regression). The lambda value that was maximized within one standard error of the lambda value yielding the lowest MSE was chosen.

4.3. Random Forest

For both our classification and regression random forest models, we used 100 trees. The importance measure used for both fitted model types was Gini impurity.

4.3.1. Classification

The number of randomly selected predictors for each loan type classification model was seven, selected via 5-fold cross-validation using Gini impurity as the measure of loss.

4.3.2. Regression

The number of randomly selected predictors for each loan type classification model was two, selected via 5-fold cross-validation using variance as the measure of loss.

5. Findings/Outputs: Results

5.1. Elastic Net Regression

The different alpha values for the elastic net regression models yielded nearly identical errors for all loan types. Because of this, we decided to choose lasso regression, to allow unnecessary variables to drop out of the model and yield the most parsimonious solution.

Loan Type	Lambda	MSE
Real Estate	0.006	0.177
CRE	0.383	159.600
Agriculture	0.020	2.805

Figure 1. Lasso regression chosen lambda values

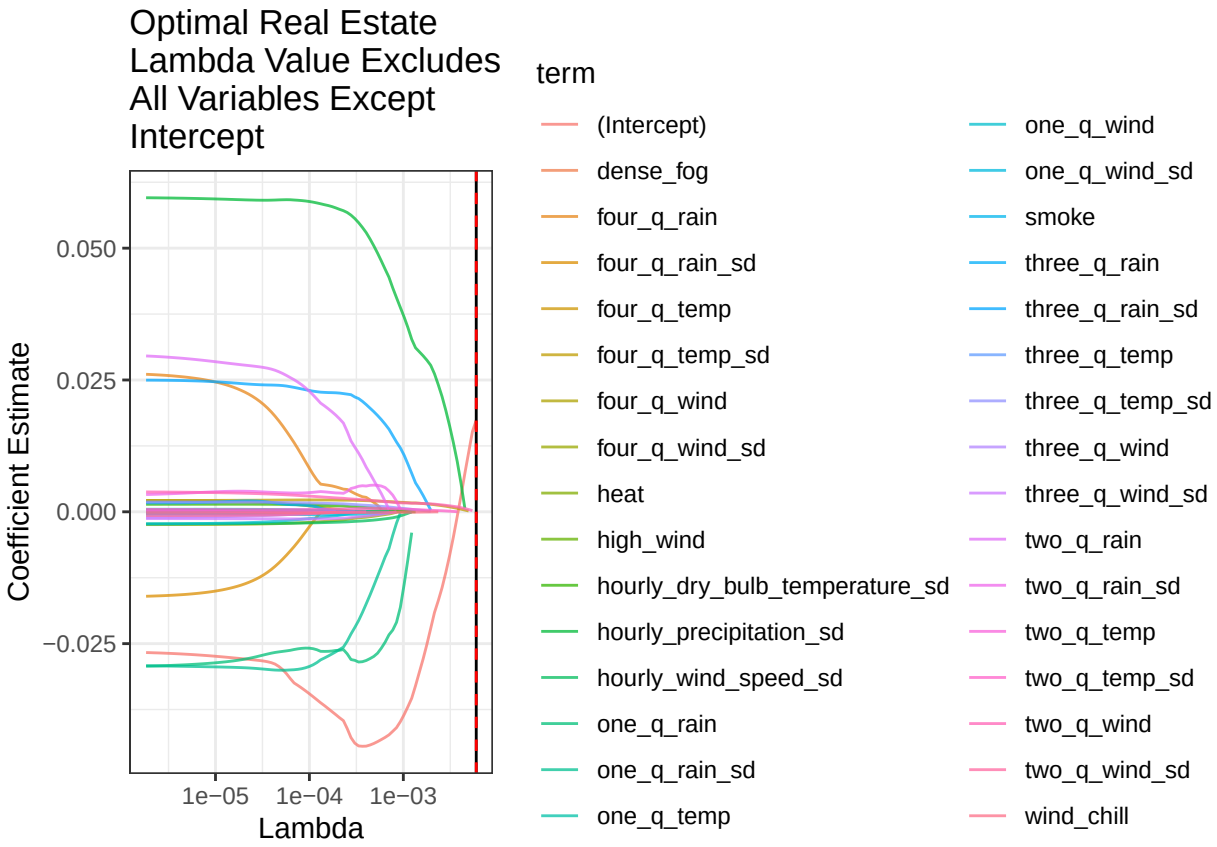

Figure 2. Real estate lasso regression variable inclusion plot

The red dotted line in Figure 2 marking the optimal lambda value shows that only the intercept is included in the final real estate model. The black line marks the lambda value that yields the lowest MSE.

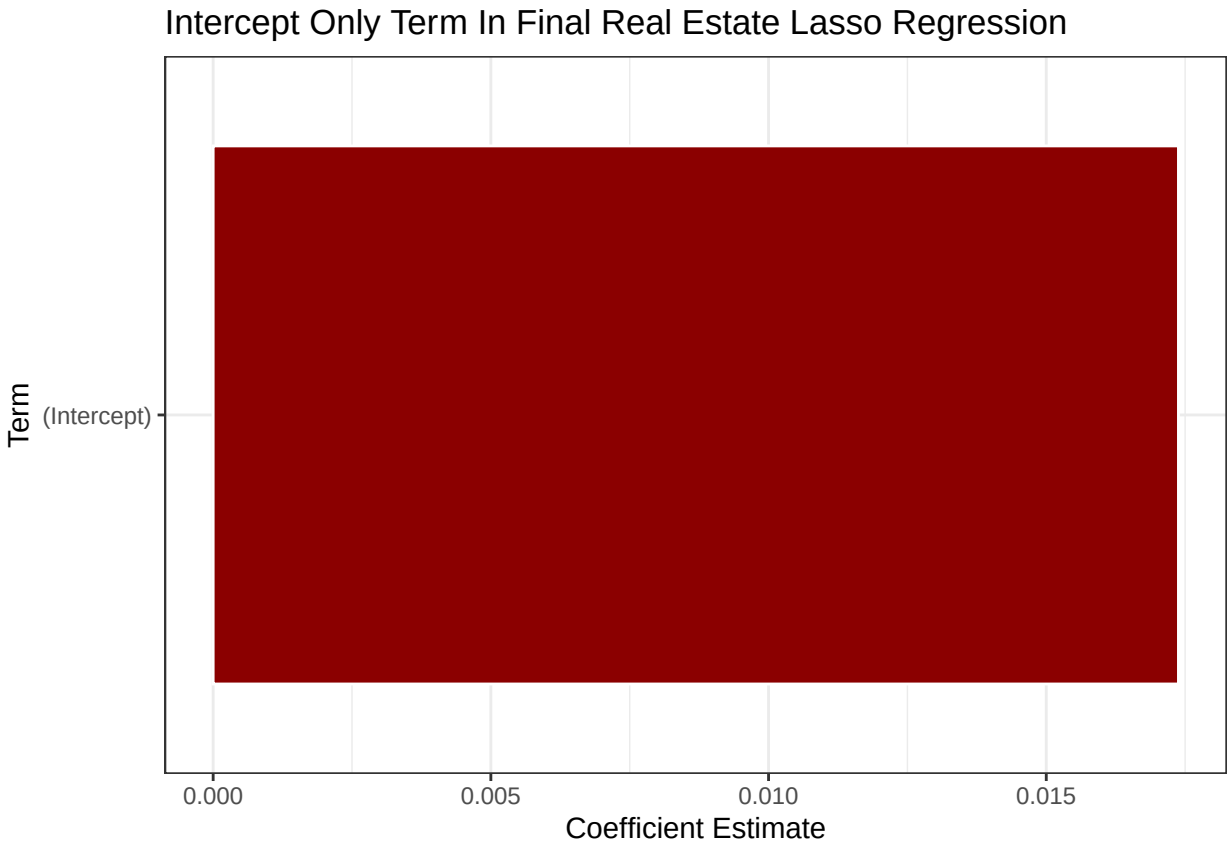

Figure 3. Real estate lasso regression coefficient estimates plot

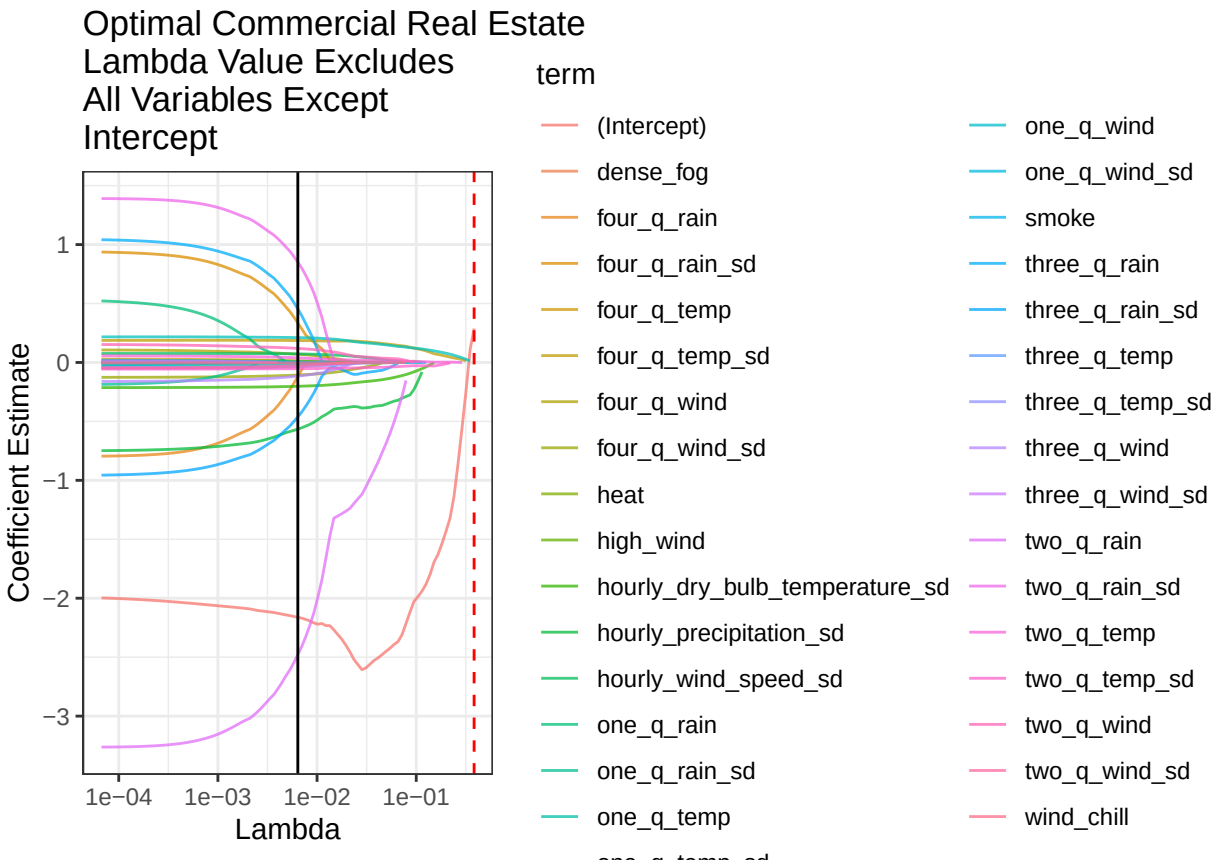

Figure 4. Commercial real estate lasso regression variable inclusion plot

The red dotted line in Figure 4 marking the optimal lambda value shows that only the intercept is included in the final commercial real estate model.The black line marks the lambda value that yields the lowest MSE.

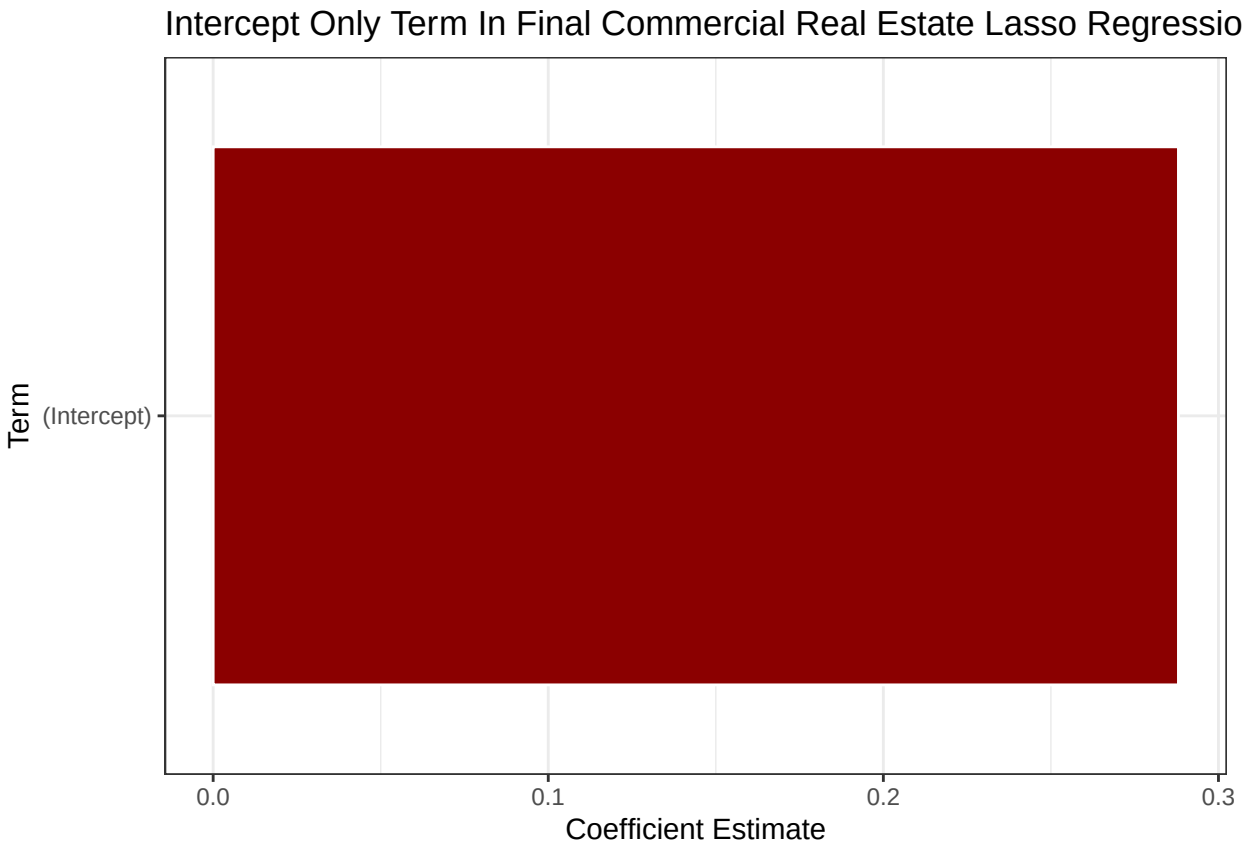

Figure 5. Commercial real estate lasso regression coefficient estimates plot

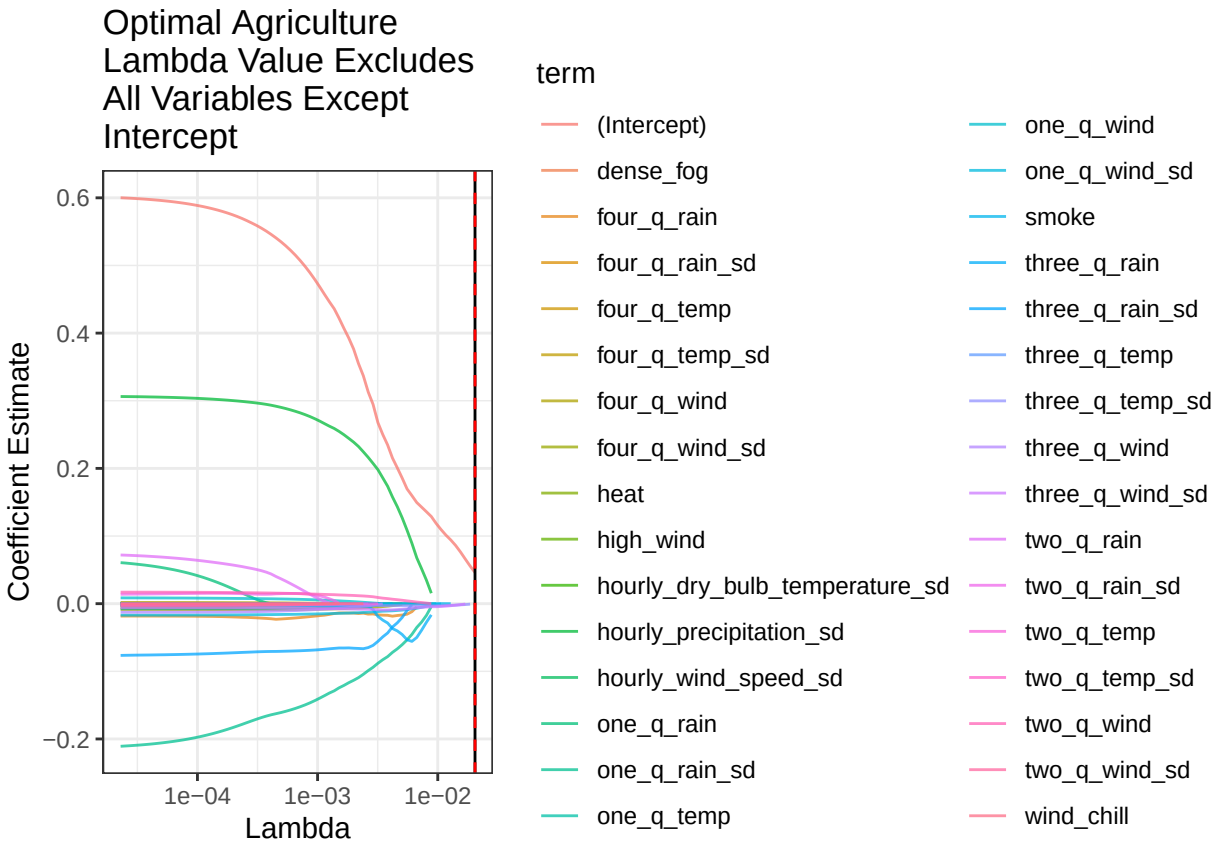

Figure 6. Agriculture lasso regression variable inclusion plot

The red dotted line in Figure 6 marking the optimal lambda value shows that only the intercept is included in the final agriculture model. The black line marks the lambda value that yields the lowest MSE.

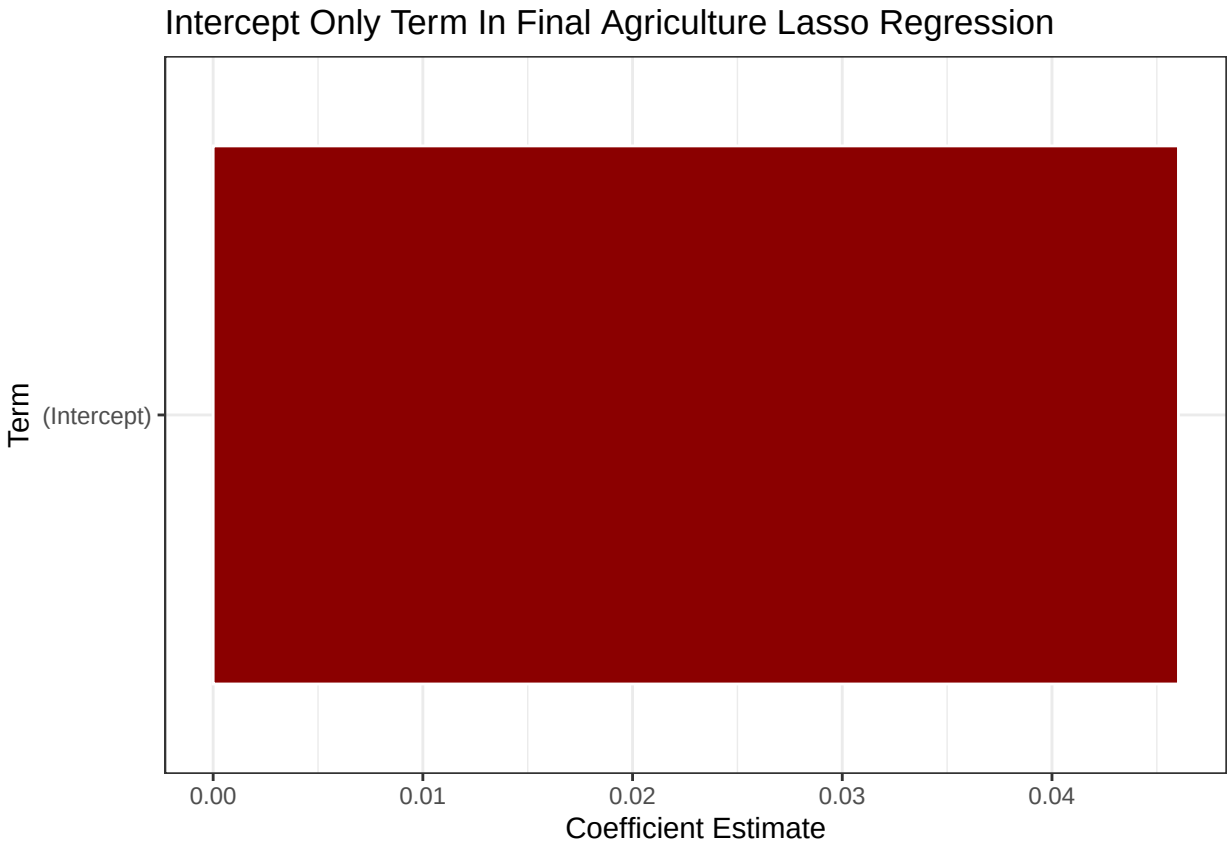

Figure 7. Agriculture lasso regression coefficient estimates plot

5.2. Random Forest

5.2.1. Classification

Loan Type	OOB Error (%)
Real Estate	46.43
CRE	46.58
Agriculture	45.42

Figure 8. Random forest classification error rates

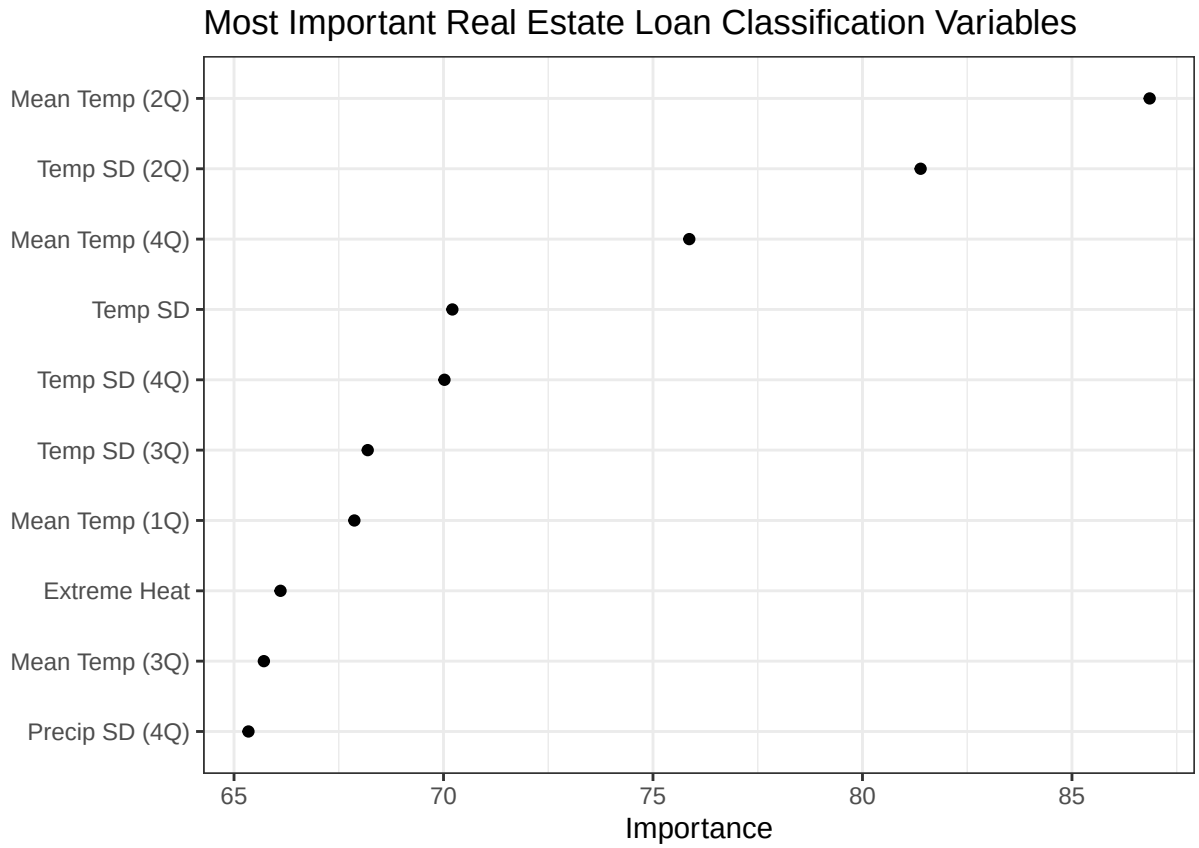

Figure 9. Real estate random forest classification most important variables

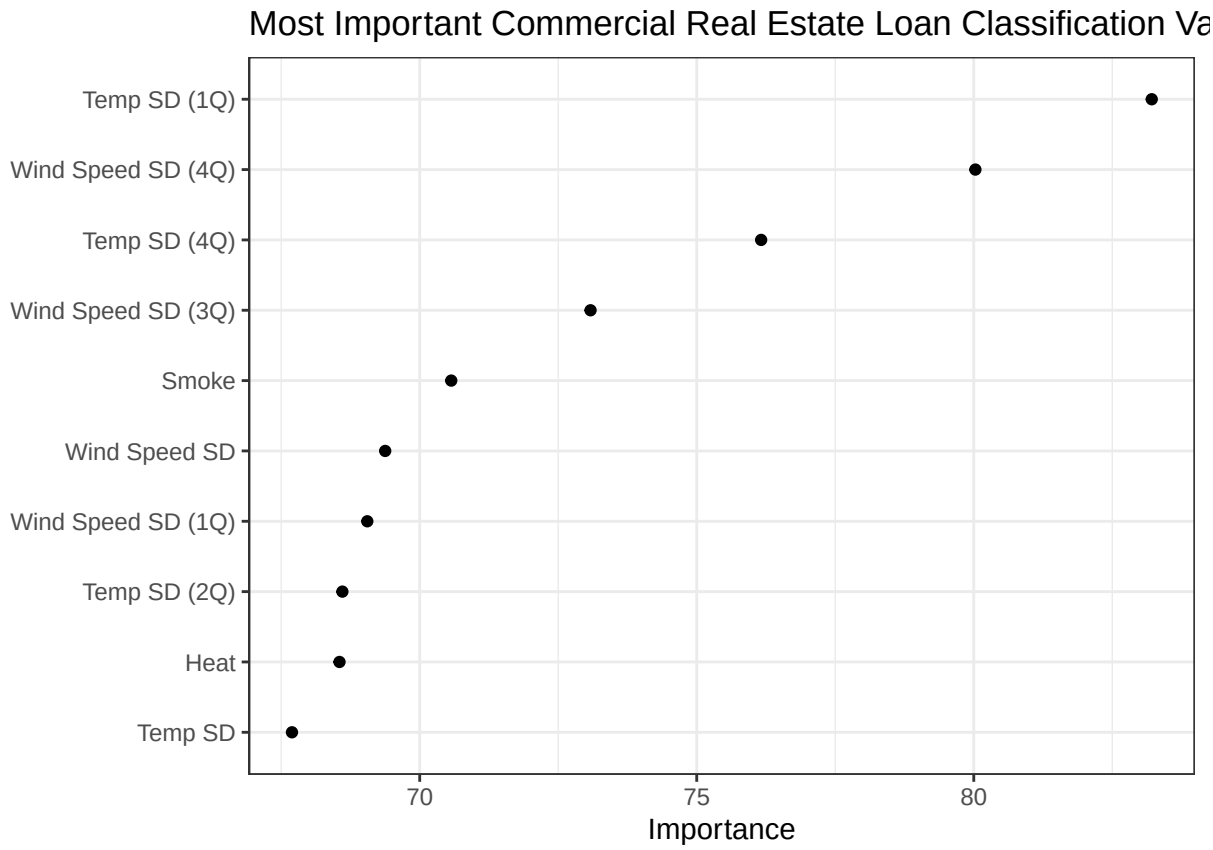

Figure 10. Commercial real estate random forest classification most important variables

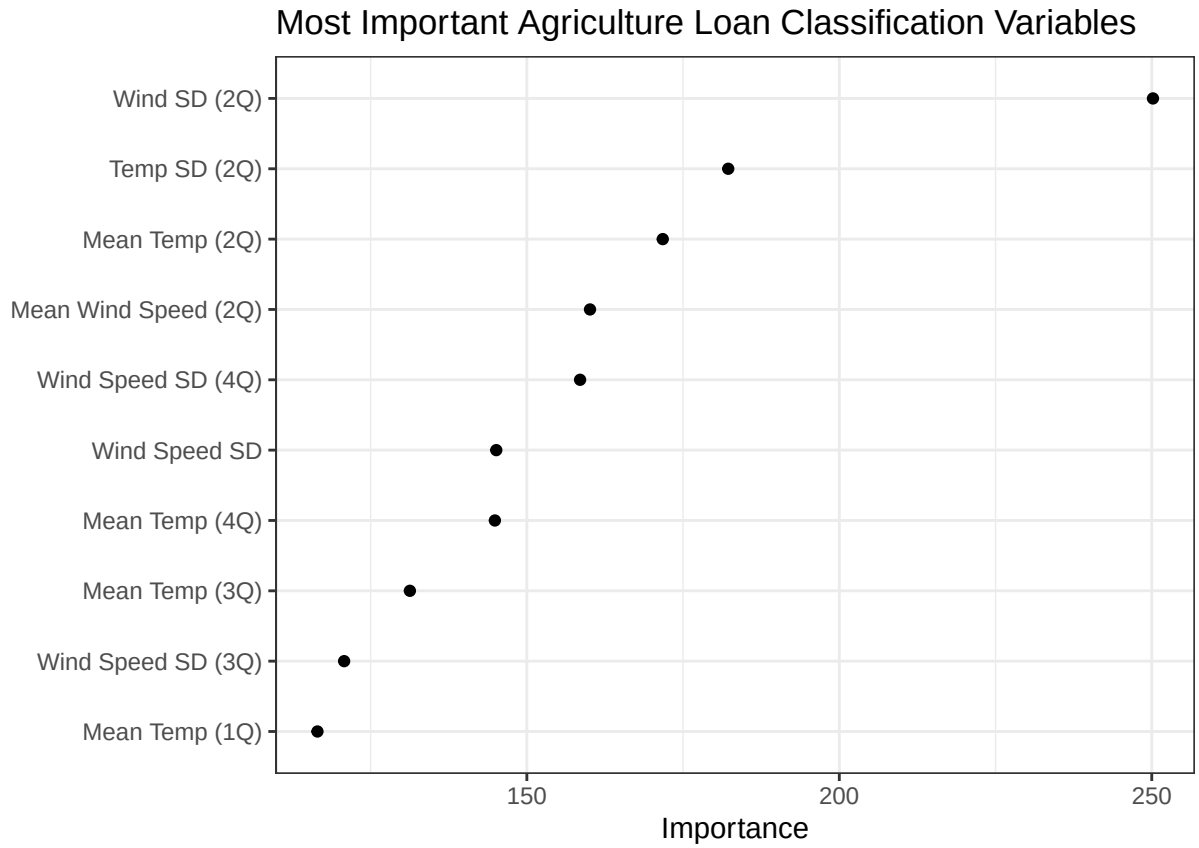

Figure 11. Agriculture random forest classification most important variables

5.2.2. Regression

Loan Type	OOB Error (MSE)	R^2
Real Estate	0.180	-0.06
CRE	167.950	-0.05
Agriculture	2.805	-0.04

Figure 12. Random forest regression error rates and R^2 values

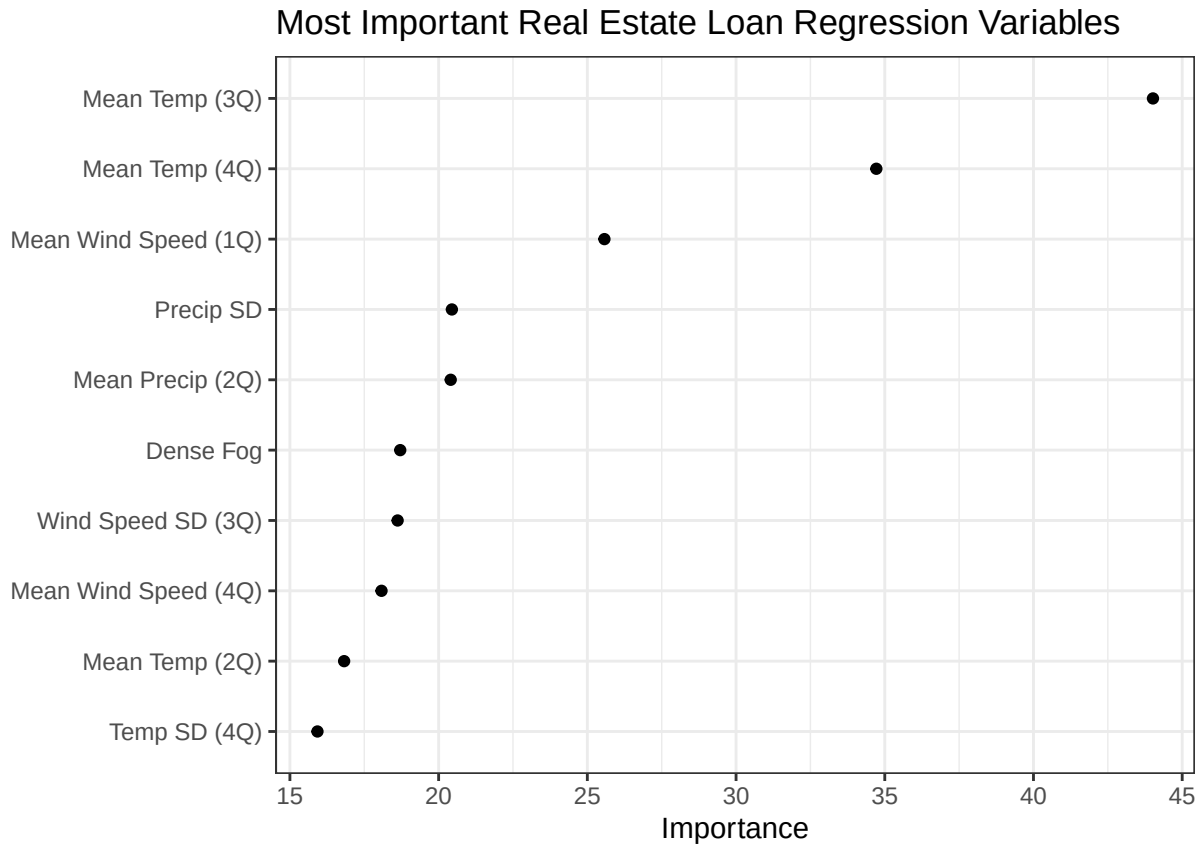

Figure 13. Real estate random forest regression most important variables

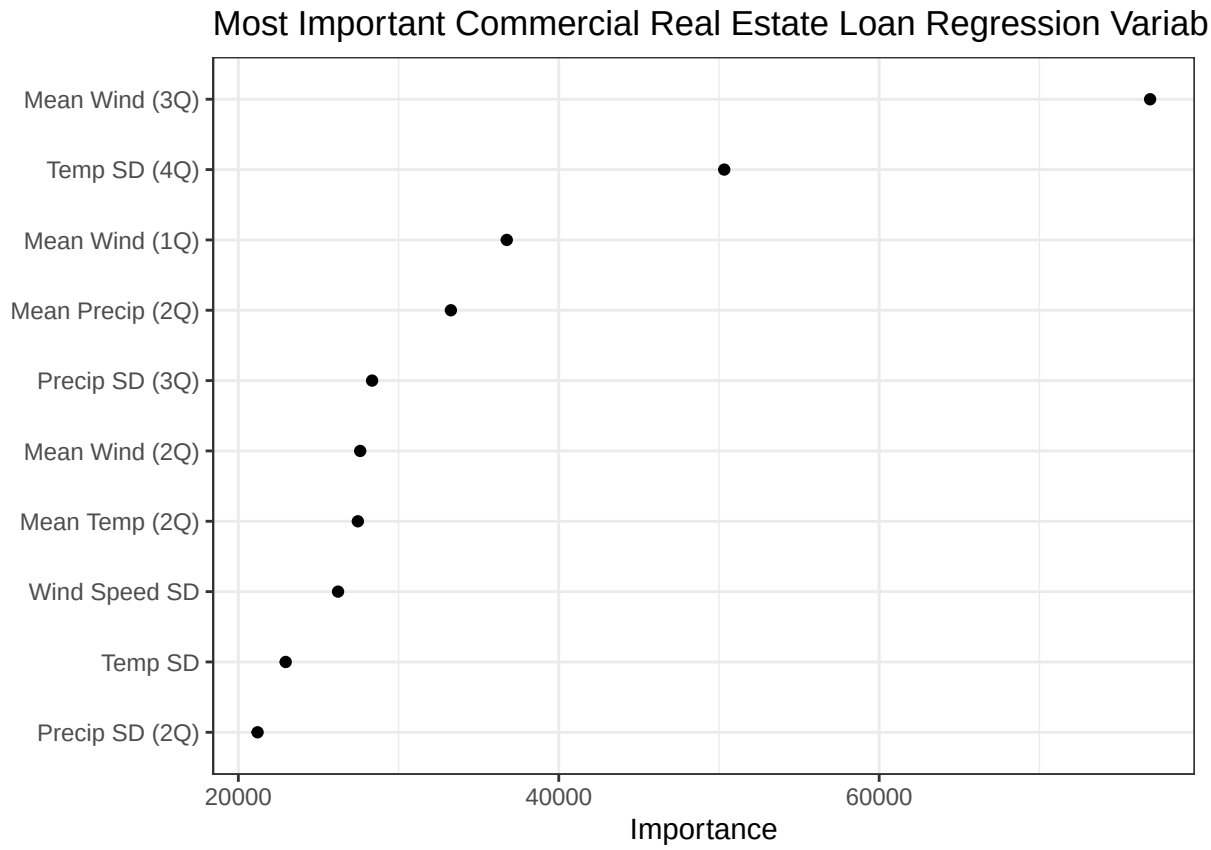

Figure 14. Commercial real estate random forest regression most important variables

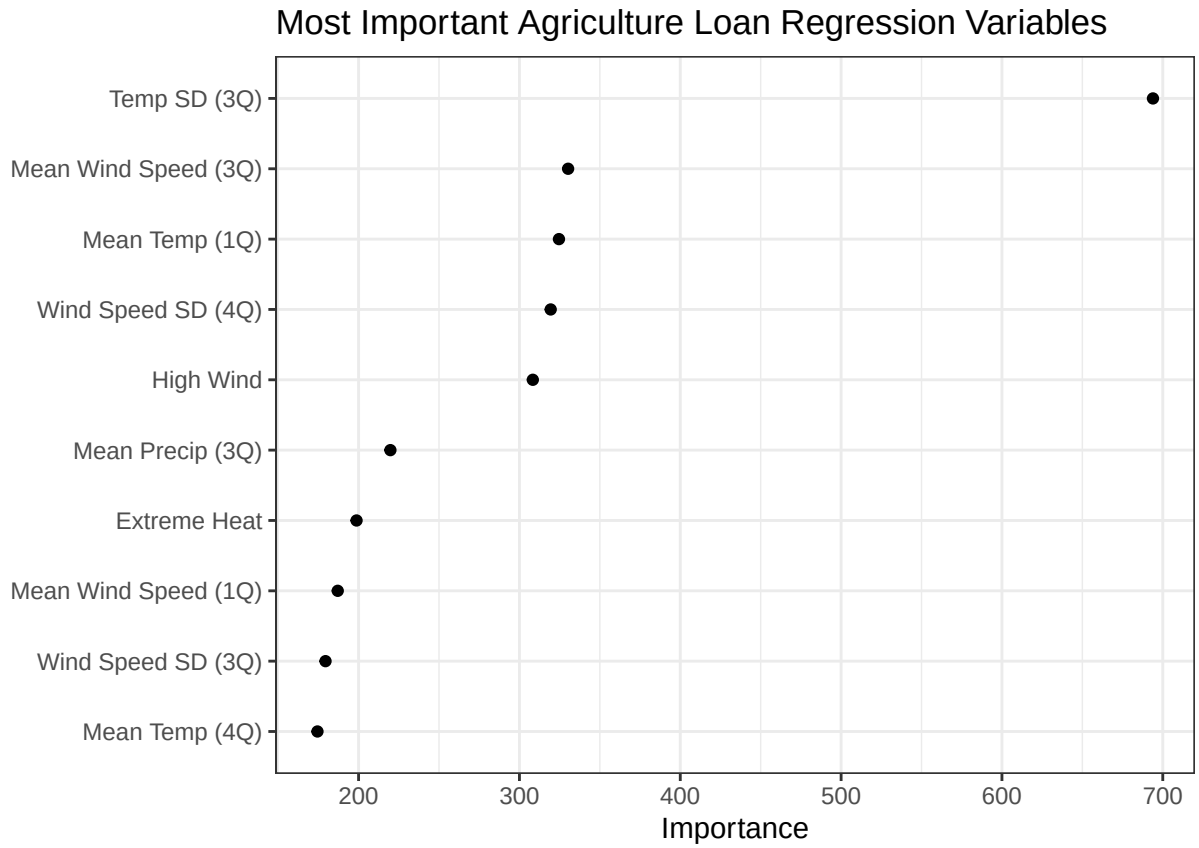

Figure 15. Agriculture random forest regression most important variables

\

The lasso regressions and random forest models all point to the same conclusion. Given the the data and our cleaning and matching methodology, none of the weather variables chosen offer any significant explanatory power for changes in real estate, commercial real estate, or agriculture lending over time. The lasso and random forest regressions all have negative R^2 values (meaning they are less predictive than the intercept alone) as seen in Figures 1 and 12, and the random forest classification achieves less than 50% accuracy (Figure 8).

6. Finding/Outputs: Discussion

6.1. Limitations and Future Work

Though our models do not point to a significant relationship between our predictors and response, the models are only as strong as the data we put in. Unfortunately we were using a sparse list of stations and some banks were matched with weather stations that were over 100 miles away. Future work could investigate pulling data from more stations that are at least within a 50 mile radius of our chosen banks and/or interpolating weather in specific areas by looking at surrounding stations. We additionally suspect there is some multicollinearity in our model, but unfortunately due to time constraints we did not investigate this thoroughly. One solution to combat this issue is to thoroughly investigate and take out correlated variables. Another solution that might solve the problem could be creating separate models for each lag length (eg 1Q lagged variables model, 2Q, etc). Lastly, we did not investigate class imbalances thoroughly in our random forest classification model, but it is possible this is an issue and should be addressed in the future via resampling techniques.

Future work could investigate different variable constructions as well. For example, our response could be constructed as loaning Z-scores or loaning percentiles compared to all banks. This would make loaning more relative instead of absolute, possibly resulting in better predictive accuracy. We

could also experiment with reconstructions of our independent variables. For example, most banks would steer clear from an area that has had consistent historic increased flooding, drought, etc, but may not be factoring in weather from, for example, one quarter ago when making lending decisions. Incorporating weather and loaning trends as a consistent time series would offer another and perhaps more realistic perspective on this issue. Additionally, aggregating up and pooling loans for counties might allow us to have a clearer and less noisy picture of the relationship between climate and lending. Because we cannot always find stations that are sufficiently close to specific banks, we could find and average the weather from stations in that county and assess impact on the county's lending overall.

Lastly, in the future we would like to incorporate and control for all relevant economic factors that may impact lending. For example, the state of the economy, competition between individual banks, and local socioeconomic conditions would all be relevant factors to control for when building a model that attempts to assess the impact of weather in the context of economic conditions.

7. Conclusion

Over the course of this project, we created a method in which to study the impacts of climate change on the loaning habits of banks from NOAA ASOS data and data available in the Call Report. Data from these two sources were filtered and aggregated to the same temporal units (quarters and years), combined, and used to create models estimating the ability of different weather variables in predicting changes in loaning habits.

Our project addressed the problems outlined in our problem statement (the lack of research available on the effects of climate change on banks' loaning patterns in the United States), by using weather data to identify both consistent, long-term climate change impacts (like increasing average temperatures), and event-based climate change impacts (like the frequency of severe storms), and using these as potentially predictive variables in understanding how banks change the way that they loan money in response to changes in local climate. We did not find statistically significant connections between climate and weather patterns and loaning behavior.

Given the time and available resources we had to complete this project, our work was not able to offer a complete understanding of how climate change impacts loaning practices across the United States, partially due to the limited time scale and geographic area we covered. With more time and resources, increasing the scope of this research to include more states and potentially a longer time scale would be useful in making our results both more useful and more resilient to irregularities in the data we worked with. Given that the United States is incredibly diverse socioeconomically, climatologically, and culturally, it is not possible to represent the whole country in three states. An addition of coastal states would be an especially interesting way to further this research, as coastal areas are subject to relatively unique and intense environmental stressors that are being worsened by climate change, and which we cannot accurately represent using only data from Ohio, Kansas, and Arizona.

8. Ethics Statement

Our team dealt with a somewhat unusual array of ethical issues throughout this project, as the data we worked with did not use humans as observational units. The NOAA data we used for this project came from Automated Surface Observing Systems (ASOS) stations, which exist in limited numbers across the geographic areas included in the scope of this project. As such is the case, we used these relatively few weather stations' observations to represent the weather conditions across the entire states included in our study, inherently limiting our results' validity to geographic areas within these states where local geographical features create weather conditions significantly different from and not able to be represented by the stations that do exist. This issue mostly came up in the communication of our results, where it was important to communicate the limitations of our work with our sponsors.

One of the other ethical concerns, fortunately somewhat addressed by the nature of this project, is the limited knowledge and lived experience each team member had with many of the subjects of our

project. None of us have lived or spent a significant amount of time in any of the three states included in our project, none of us have specific knowledge about loaning practices, commercial real estate, or agriculture, which became especially apparent during the early stages of the project as we waded through the potential information we could consider within the scope of our study and wrestled with certain decisions about what to include and what to leave out. One way we addressed this issue was by deferring to our sponsor in making certain decisions, as we trusted that their greater familiarity with these topics would allow more appropriate choices to be made in selecting factors such as the variables and states chosen to be included in our models and analysis. Our sponsor also stressed the importance of not allowing these gaps in our knowledge to be filled in by assumptions we made about a bank or the community it serves based on its location, size, or loan portfolio.

Were we to rely on our assumptions and limited knowledge when making project-related decisions, the conclusions we may have come to might have ended up being both inaccurate and unhelpful at best and harmful at worst. Although our project was serving more as a proof-of-concept demonstration, it is still reasonable to worry about the ethical implications of linking climate change impacts to loaning behavior, as it might further discourage certain types of loaning in impacted areas which in turn could encourage less protected or extra-legal loaning practices to people in these impacted communities who find themselves unable to secure less-risky loans from more traditional financial institutions. While ideally this impact would be avoided entirely, it is especially problematic if these practices occur due to faulty analysis based on assumptions made with little background knowledge.

Abbreviations

The following abbreviations are used in this manuscript:

CFRB	Cleveland Federal Reserve Bank
FFIEC	Federal Financial Institutions Examination Council
NOAA	National Oceanic and Atmospheric Administration
FDIC	Federal Deposit Insurance Corporation

References

1. 2013.
2. Kolbert, E. How Did Fighting Climate Change Become a Partisan Issue? *The New Yorker* **2022**. Section: comment.
3. Bezner Kerr, R.; Hasegawa, T.; Lasco, R. Food, Fibre, and Other Ecosystem Products. *Climate Change 2022: Impacts, Adaptation and Vulnerability* **2022**, pp. 713–906. doi:10.1017/9781009325844.007.
4. Javadi, S.; Masum, A.A. The impact of climate change on the cost of bank loans. *Journal of Corporate Finance* **2021**, *69*, 102019. doi:10.1016/j.jcorpfin.2021.102019.
5. FFIEC. Download Bulk Data - FFIEC Central Data Repository's Public Data Distribution.
6. NOAA. NOAA Local Climatological Data Archive.
7. 1998.
8. 2004.