April 16th, 2023

COS: Project Report

**Introduction to Project/ Background Information**

Established in 1889, Stillwater, Oklahoma, the home of Oklahoma State University, has transformed itself from a small agricultural community to a micropolitan area with a market area of 85,000 people (about the seating capacity of the Los Angeles Memorial Coliseum) during the school year and approximately 46,000 permanent residents. A unique characteristic of Stillwater is that the main power driver for its economy is the campus. However, as the City of Stillwater continues to grow, the municipality building has noticed a growing problem of people defaulting on their utility bills, which poses a financial obligation to the taxpayers of the community. This concern of delinquent account is becoming increasingly significant as the community is responsible for the debt of their neighbors and business owners.

To address this issue, a Business Analytics and Data Science year-long project has been initiated to assist the micropolitan area in predicting potential delinquent accounts, discovering risk factors in their temporary and permanent citizens, and allowing the City of Stillwater to intervene to help these individuals pay their utility bills on time. This project will provide valuable insights that can help the City intervene early on and take corrective action, reducing the burden on the taxpayers and improving the financial stability of the community.

**Business Problem/ Questions**

The City of Stillwater is faced with a significant business problem related to delinquency issues in the local community regarding utility bill payments. As households or businesses default on their utility bills, the city incurs a financial burden that accumulates and results in substantial financial losses. The responsibility of collecting delinquent accounts falls on the city, which can be a laborious and costly process. If left unaddressed, these delinquencies can negatively impact the city’s credit ability to keep the costs out of the taxpayers' pockets.

Therefore, it is essential to develop a strategy to predict and prevent delinquencies in utility bills to reduce the burden on taxpayers and improve the financial stability of the community. By leveraging data science techniques and census data, the City of Stillwater can identify patterns and correlations between the demographics of households and the likelihood of delinquency. This analysis can be used to create a predictive model that identifies households that are at higher risk of becoming delinquent on their utility bills.

Using this model, the city can intervene early and take corrective action to prevent delinquency and reduce the burden on taxpayers. Additionally, this strategy can help the city improve the quality of life in the community. By addressing this business problem proactively, the City of Stillwater can ensure the financial stability of the community and provide better services to its residents.

The questions we want to take on include:

* Understanding the **data library –** This includes understanding all the variables form the various tables provided, understanding the format, metrics to quantify, and the relation of the variables to our broad question of the effects of those variables on delinquency.
* Observing **clusters and trends** followed by accounts – This includes plotting graphs to observe or identify patterns to choose specific variables to further investigate their impact on delinquency.
  + How much revenue is lost due to delinquent utility bills, and what is the impact on the City’s budget? How does this affect the citizens of Stillwater?

* Relating said patterns to **delinquent accounts –** Understanding the impact of the variables on each other and checking for correlation among them.
  + Are there any external factors, such as economic conditions, that contribute to delinquency, and how can they be mitigated?

* Creating model to **predict the credit offense amount** for new and existing users – To create different models on the variables selected and comparing their accuracies.
  + How can the city measure the success of its strategy to prevent delinquency and what metrics should be used to track progress?

As we delved into the data and conducted our analysis, we aimed to answer several key questions to gain deeper insights into the delinquency issues related to utility bill payments in the City of Stillwater.

Through our research, we hoped to gain a better understanding of the factors that contributed to delinquency risk and determined the most effective strategies to mitigate delinquency.

Furthermore, we explored the demographics of households that were more likely to become delinquent on their utility bills and assessed how these factors could be leveraged to prevent delinquency.

By answering these questions and gaining valuable insights into the data, we developed a comprehensive and data-driven approach to address the delinquency issues in the Cit of Stillwater.

**Data description and Data Collection:**

The Data is provided in 4 different tables which contain various elements involving the customer details and their billing information. We have four tables containing about 61 variables, each of which gives us information about the location of the customer, their deposit details, or their bills. However, some of these tables and variables were null, and we are ignoring those variables for our analysis.

PFB the data dictionary for the key tables which we are using in our analysis:

* UTCSID – Customer ID
* UTLCID – Location ID
* SVC – Type of Service
* NewDep – Deposit Amount
* Date – Deposit Date
* Address – Address of connection

New Deposits (New\_deps)

* UTCSID – Customer ID
* UTLCID – Location ID
* Credit Offense Type – Credit Offense Type
* Credit Offense Date – Credit Offense Day, Month, Year

CREDIT HISTORY (UT420AP)

|  |  |
| --- | --- |
| **New Deposits (New\_deps)** | |
| **Column Name** | **Description** |
| UTCSID | Customer ID |
| UTLCID | Location ID |
| NewDep | Deposit Amount |
| Address | Address of connection |

|  |  |
| --- | --- |
| **CREDIT HISTORY (UT420AP)** | |
| **Column Name** | **Description** |
| UTCSID | Customer ID |
| UTLCID | Location ID |
| Credit Offense Type | Credit Offense Type |
| Credit Offense Date | Credit Offense Day, Month, Year |
| Credit Offense Date | Credit Offense Day, Month, Year |

* **UTCSID** and **UTLCID** are the primary keys which help to identify a unique customer.

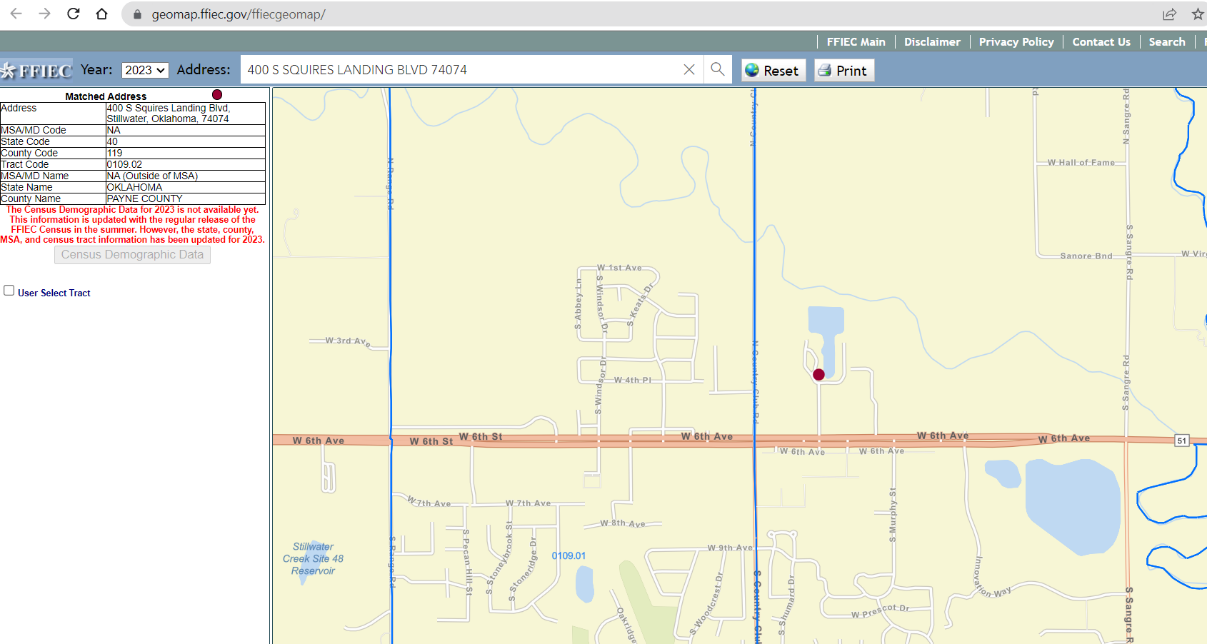
After removing all the commercial customers from the dataset, we have following data facts which were considered for our analysis and prediction model:

* Total number of distinct residential UTCSID and UTLCID: **16,516**
* Distinct residential addresses where City of Stillwater provides services: **6,918**

**Scrapping data:**

As part of the secondary research, we were planning to add socioeconomic data to our existing company data set. To be able to get socioeconomic parameters, we are looking at census data. However, we have census data on an aggregate level of a tract rather than individual addresses. Therefore, to accommodate that we have worked on finding the corresponding tract values for individual addresses.

We have found the website: '<https://geomap.ffiec.gov/ffiecgeomap/>' This website outputs the tract for any address given as an input. Since we had over 4000+ individual rows, we decided on writing a scraping code instead of manually entering the addresses to the site.



Following are some of the important steps we completed for the web scrapper

* + Firstly, the scrapper would load the given URL of the web page
  + Then the scrapper would read every row from the given column and place it in the search bar
  + The scrapper would then wait for a certain time for the results to load
  + When the results are loaded, the scraper will copy the tract output from the output section and store it in a list.
  + The scraper then clears the search section and pastes the next row as an input.
  + This process is continued for all the rows in the individual addresses.
  + All the output tracts are stored as a list and saved to the local address as a column for a CSV

**Data Preparation:**

To prepare the dataset for our analysis, we followed the below steps:

1. We used the credit table to extract the customers who have defaulted at least once.
2. To remove the duplications in the credit table, we kept the most recent credit amount for each combination of UTCSID and UTLCID (Cust\_id).
3. To include just the residential customers, we merged the modified credit table with the list of residential UTLCID and their corresponding addresses.
4. Now, we had the dataset with UTCSID, UTLCID, most recent credit amount, credit default date and the full address.
5. To enrich the dataset, we used the scrapped census data and merged them with corresponding census tracts based on the full address.

At the end of the above steps, we collated the census data with our modified credit dataset to create the final dataset which include the following variables:

|  |  |  |
| --- | --- | --- |
| **Source** | **Variable Name** | **Description** |
| City of Stillwater | utcsid | Service ID |
| City of Stillwater | utlcid | Location ID |
| City of Stillwater | cust\_id | Unique ID based on combination of utcsid and utlcid |
| City of Stillwater | credit\_date | Most recent date of credit default |
| City of Stillwater | credit\_offense\_amount | Most recent credit amount defaulted |
| City of Stillwater | full\_address | Complete address of the customer |
| Census | census\_tract | The census tract under which the customer falls into |
| Census | avg\_age | Average age of the population in the census tract |
| Census | population | Average populatioon in the census tract |
| Census | male | Male population in the census tract |
| Census | female | Female population in the census tract |
| Census | median\_income | Median income of the population in the census tract |
| Census | owners | Number of owners in the census tract |
| Census | renters | Number of renters in the census tract |
| Census | total\_housing | Total number of housings in the census tract |
| Census | underserved | Yes or no based on if the census tract falls under underserved category |
| Census | income\_level | Income level of the population in the census tract |

**Census Data and Demographics:**

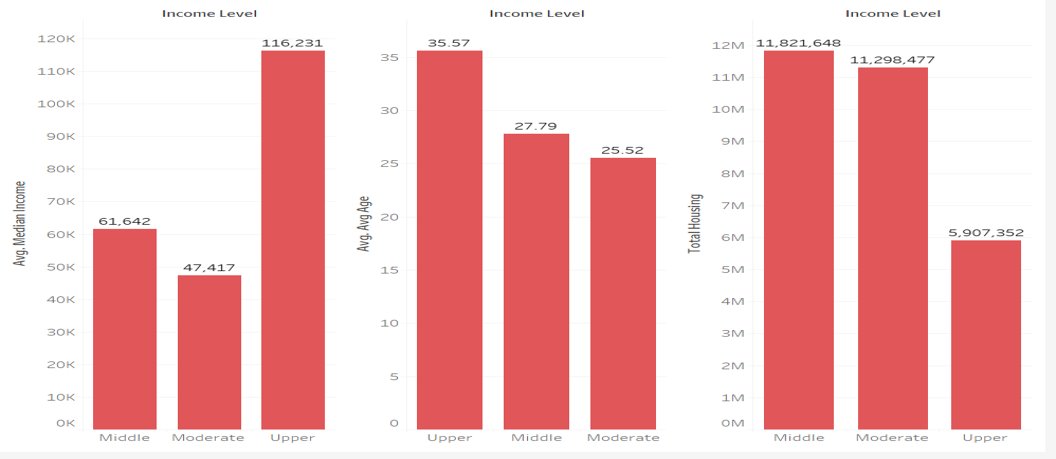
In the pivotal predictive phase of our year-long project, we employed a range of essential tools and resources, including Python, Excel, Tableau, and census data, to further refine our analysis. Census tracts proved to be an asset in augmenting our understanding of the City of Stillwater. Building on our prior findings, we conducted a thorough examination of the census data to uncover invaluable insights into the community's demographics, socioeconomic status, and housing situation. To conduct a detailed exploration of this data, we employed local community census blocks, which enabled us to perform a more comprehensive analysis of the city and its data. This approach allowed us to not only identify general demographic information such as age, gender, and race but also more specific characteristics like household income.

**Methodology:**

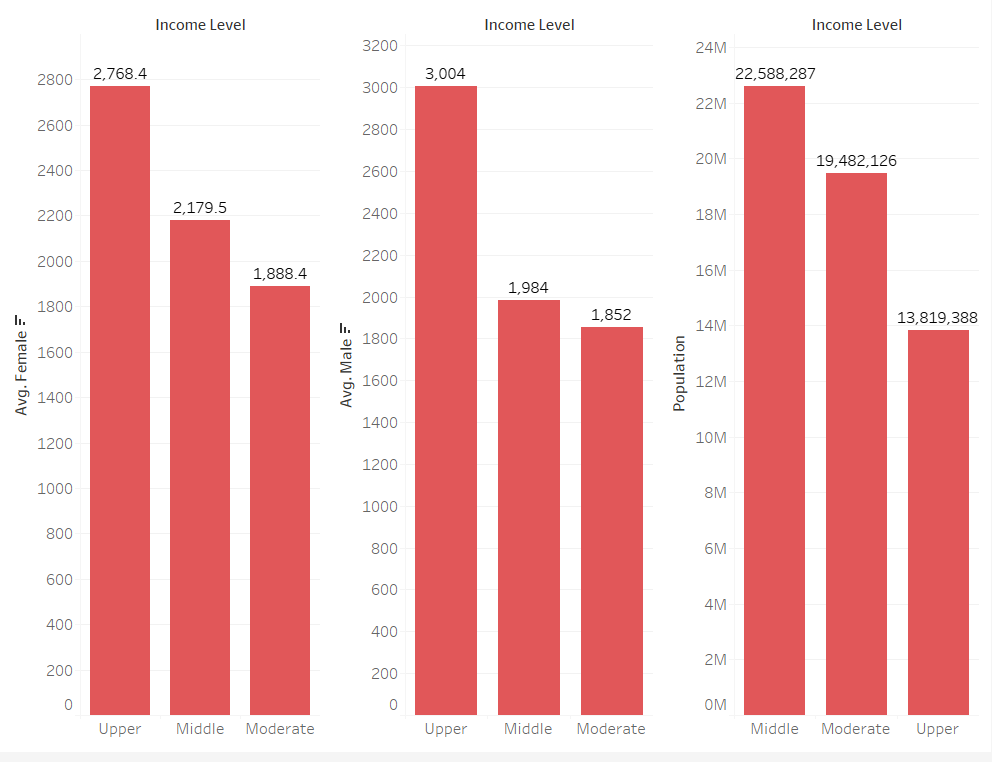
During the last semester, we examined how different areas within the Stillwater community behave in terms of credit owed for various utilities. We accomplished this by matching addresses to corresponding census tracts and augmenting the data frame with socio-economic data obtained from the census bureau. Using this finalized data frame, we then developed predictive models to identify the factors that influence credit amounts. This empirical analysis aims to establish causal relationships. To better understand the data, we created visualizations, and the resulting models and insights from the visualizations are presented in our results.

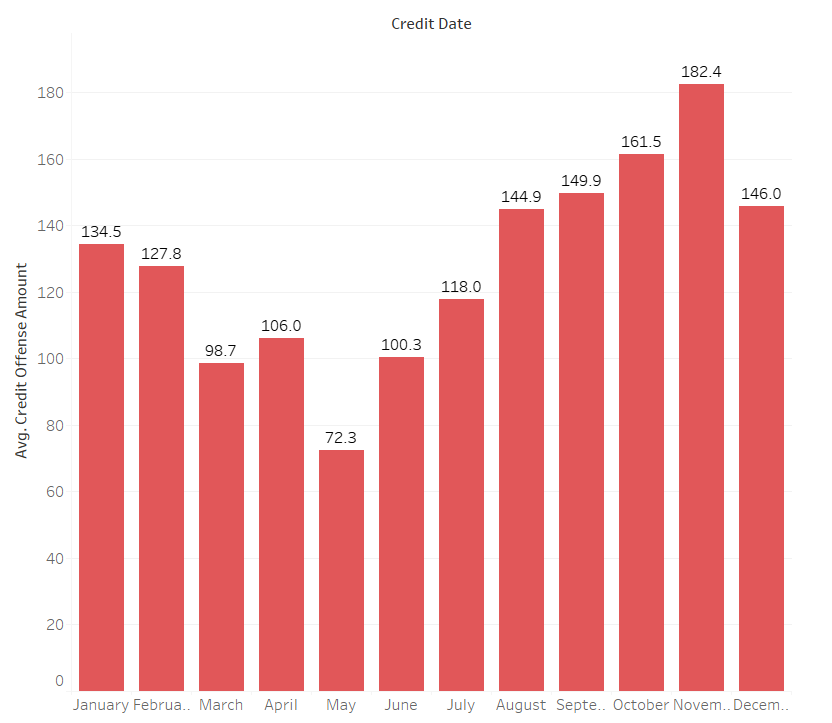
**Results:**

We carried out some basic visualizations to better understand the dataset's descriptive statistics.

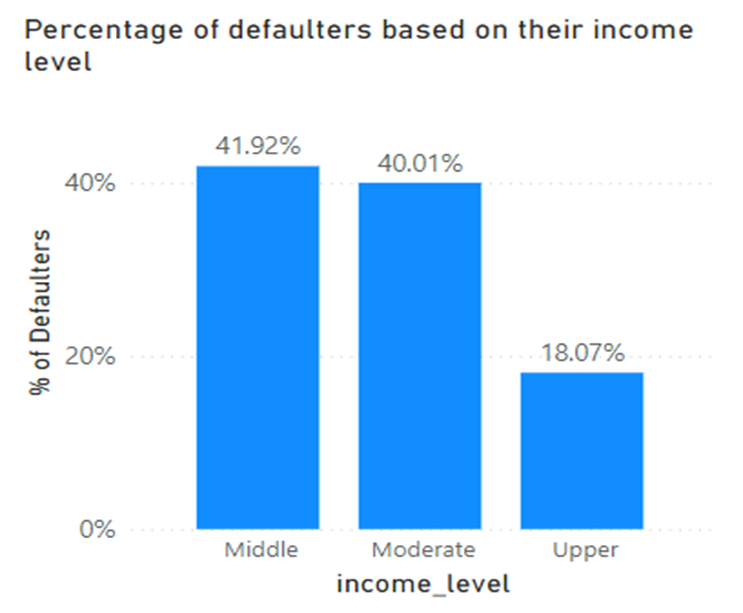
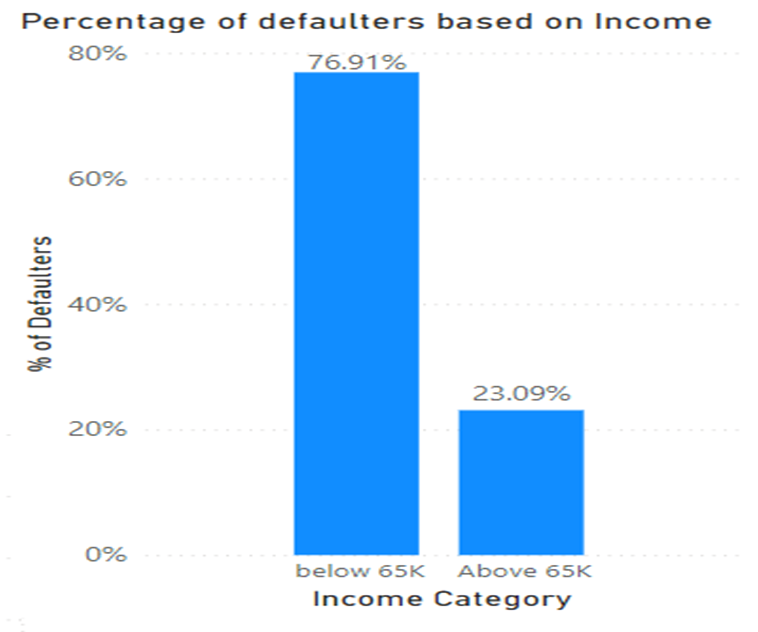


Visualizations that showed the distribution of income across various categories showed that the average income by income levels showed people in the upper-income level who had an average age of 35.67(36) had an average annual income of $116.231. Ignoring the null values, total housing for people in the upper-income level was the lowest among other income levels.

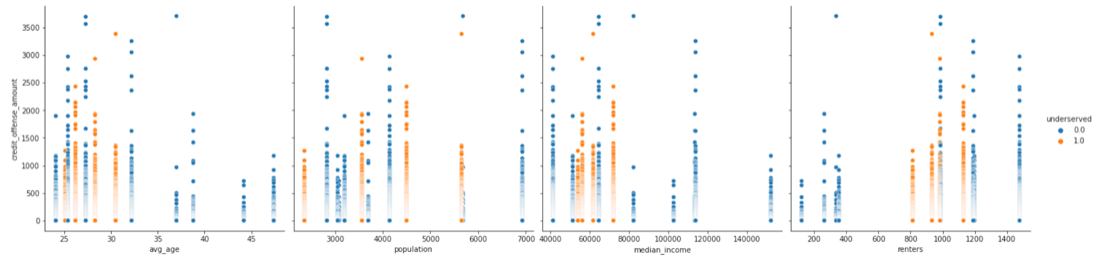
  
The income-level by gender population visualizations showed that majority of men and women were in the upper income-level but a population against income-level chart showed that majority of the population belonged in the middle-income level.

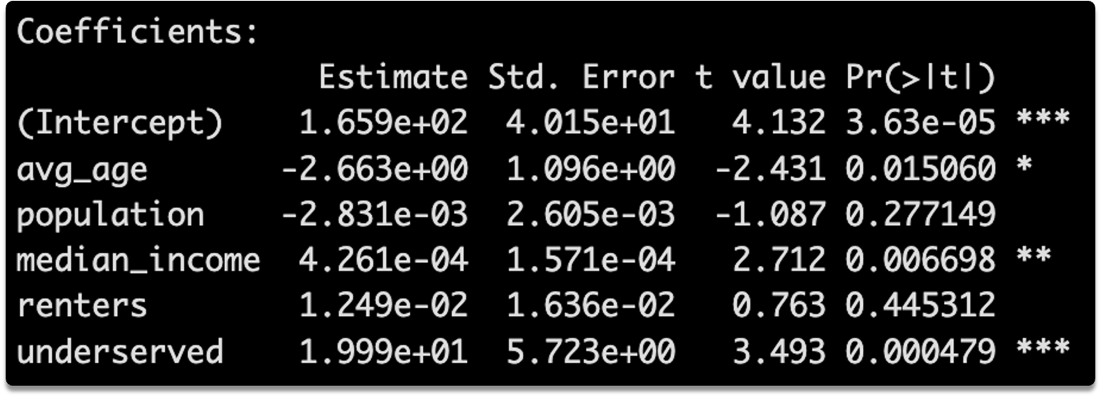


The analysis also showed that November had the highest amount of credit offence followed by October and September. The trend in the credit offence amount over the years gives the assumption that people tend to accrue more utilities during ember months and the beginning of the year.



The shows that about 76% of the customers who defaulted earned less than 65k per annum and that customers that belong to the middle-income level defaulted the most. This is closely followed by customers who belonged to the moderate income-level.

We did some secondary research to understand what kind of socioeconomic factor might be related to the amount of money that people are defaulting on. Therefore, we plotted a scatterplot to understand the kind of relationship that exists between the target variable and the predictor variables.   
  
  
  
We see that none of our predictor variables (“avg\_age”, “population”, “median income”, and “renters”) have any kind of identifiable relationship with our target variable (“credit\_offense\_amount”). This might be due to the fact the values of our predictor variables are only a few based on the census tract (only 12 different values for each of our predictor variables) and our target variable’s range is 3707.92 with a mean of 128.551 and 8687 unique values.  
Even we understand that there is not an identifiable linear association between the target and predictor variables, we look at the correlation matrix to understand the strength of linear association.   
  
  
  
We notice that none of the predictor variables have a strong correlation with the target variable.

However, we have gone ahead and fitted a linear regression to understand the significance and the kind of association that the predictor variables might have with the target variable. The regression model has a value of 0.2% which means that around 0.2% of the variation in “credit\_offense\_amount” is being explained with the variables “avg\_age”, “population”, “median\_income”, “renters”, and “underserved”. The regression model has been checked for multicollinearity and the heteroscedasticity and adjusted for the same.   
  
  
  
We see that the amount of credit offense decreases with the increase in the average age of that particular census tract. Since Stillwater is predominantly a college town, we can infer that the student population is defaulting more than the regular residents of the city. We also notice that the census tracts with the higher median income are incurring higher credit amounts. This might be because the residents in wealthier neighborhoods might have more opportunity to take on debt or the apartment complexes that they live in are an all-bills paid community and the community itself might incur the credit offense amount. We also see that the underserved census tracts have a higher credit offense amount as compared to the other census tracts.  
  
Following is a table containing model comparison of all the models that we have built.

|  |  |  |
| --- | --- | --- |
| Models | Root Mean   Squared Error | R-squared (1 – TSS/RSS) |
| Boosting | 223.0759 | 0.0106 |
| Bagging | 223.0874 | 0.0105 |
| Random Forest | 223.1019 | 0.0104 |
| Decision Tree | 223.1951 | 0.0096 |
| Linear Regression | 223.9551 | 0.0028 |

We see that Boosting is the champion model over here. The root mean squared error value is the least and the value is the most for this particular model.   
We can try to fit other models and check whether those models perform better. But these are all black box models that make it very difficult to understand how the predictor variables are associated with the target variable.

**Discussion/Conclusions:**

For this project, we used the data available in the census bureau to append to the existing credit offense amount. This analysis gave us insight into how the different addresses in still are affecting the delinquency rate of the city.

The next steps in the data analysis include but are not limited to:

* Further visualizations that can help uncover additional insights and uncover potential business issues for utility companies will be carried out. Additional data about the customers will be required to build out a personalized predictive model for the citizens of Stillwater.
* Data from the smart meters (that are going to be installed in the later part of 2023) can help build better predictive models for understanding the credit offense type. We can create an alert system using the existing predictive models to make sure that the residents are not defaulting.
* Entire data about all the residents with a flag about whether they have defaulted in the past if given to us can be fitted into a model to understand whether a person taking a new connection will be defaulting.

**References - Team**

Utility & Billing Services. (n.d.). Retrieved November 7, 2022, from <http://stillwater.org/page/home/government/utilities/utility-billing-services>

How utilities can use machine learning for bad debt control. (n.d.). Retrieved November 7, 2022, from <https://articles.utegration.com/how-utilities-can-use-machine-learning-for-bad-debt-control>

Stillwater, Oklahoma population 2022. (n.d.). Retrieved November 7, 2022, from <https://worldpopulationreview.com/us-cities/stillwater-ok-population>

Donadello, I., Hunter, A., Teso, S., & Dragoni, M. (2021, December 15). Machine learning for utility prediction in argument-based Computational Persuasion. Retrieved November 7, 2022, from <https://arxiv.org/abs/2112.04953>

City Of Stillwater Interactive Map. (n.d.). Retrieved November 7, 2022, from <https://stw.maps.arcgis.com/apps/webappviewer/index.html?id=e5f507cfcbe445d5a5b9277ab706a2bc>

**Appendix:**

The following are the Tableau workbooks, Python and R codebooks that were used to create the models and the visualizations. We have hyperlinked all our workbooks with the titles.

[Data Preparation](https://ostatemailokstate-my.sharepoint.com/:u:/r/personal/shreyan_datta_chakraborty_okstate_edu/Documents/Study%20Materials/2nd%20Semester/BAN%205560/CP/Codes/CP_Ranjith_Final_working.ipynb?csf=1&web=1&e=wOWJu0)[Web Scraping 1](https://ostatemailokstate-my.sharepoint.com/:u:/g/personal/shreyan_datta_chakraborty_okstate_edu/EXfgS_dedaJNicDBlLnr3I4BiYBaBLwnMZA-F4vn7c_AxQ?e=wFepcR)

[Web Scraping 2](https://ostatemailokstate-my.sharepoint.com/:u:/g/personal/shreyan_datta_chakraborty_okstate_edu/EYA2i1YnWFhEphFucUpDccMBgeQgkhBL-EOvBur3LGhtXQ?e=vUCexr)

[Web Scraping 3](https://ostatemailokstate-my.sharepoint.com/:u:/g/personal/shreyan_datta_chakraborty_okstate_edu/EUH8Nadk6iVKvPXbezxCdR4B4JClTKI0d3cq8Oue7l4prg?e=AyNSRb)

[Modelling 1](https://ostatemailokstate-my.sharepoint.com/:u:/g/personal/shreyan_datta_chakraborty_okstate_edu/Ef1bMFscRUZHnPzNfWwVjNIBZFYpdhF_5t2IbeMhpuDe1Q?e=Xd7CmI)

[Modelling 2](https://ostatemailokstate-my.sharepoint.com/:u:/g/personal/shreyan_datta_chakraborty_okstate_edu/EZZhQt2vBEdIsxmG7-U97coBQHIUNheLk44lfJn73a3U7A?e=AIsoWw)

[Tableau Workbook 1](https://ostatemailokstate-my.sharepoint.com/:u:/g/personal/shreyan_datta_chakraborty_okstate_edu/EQL9paTaexBNoh6C9ZAPCIsBdPKI2cs2ScbjNyz8DtPVDg?e=RO2rRl)