AI Prediction of Michigan Water Use

Jacob Javier

Springboard Data Science Career Track

**AI Prediction of Michigan Water Use**

With the human population still increasing, the demand for consumer goods is increasing as well. To keep up with demand, natural resources like water are heavily impacted to produce goods and services. For example, the Great Lakes provides a source of revenue for recreational businesses, helps generate hydroelectric energy, and is used to in industrial manufacturing. Although water levels have been on a steady incline since 2013; there was a 73% decrease in water levels from 2020 to 2023 (U.S. Environmental Protection Agency 2023).

In June 2024, Martusiuk aggregated the Department of Environment, Great Lakes, and Energy’s water usage data for the Great Lakes Basin from 2013 to 2022. This decade’s worth of data can be used to gain insight into the increasing demands on Michigan’s water supply using regression analysis and forecasting.

The primary objective of this project is predicting the future impact on Michigan’s water supply based on archived industry and geographical water consumption levels. More specifically, this project will showcase how water consumption changes over time across industries. Population growth data will also be included as an explanatory variable for the increased demand on the industries. With the annual population data, correlations may be identified across the industries to explain changes in demand. Additionally, highlighting the relationships between industries’ water consumption will provide insight into any underlying effects on water use.

**Data Wrangling**

The water use data set (2024) originally contained eight features and 6630 observations. The data ranged from 2013 to 2022 and showed the water use data of seven unique industries that was sourced three different water sources. The seven industries that were evaluated were Commercial-Institutional, Electric Power Generation, Irrigation, Livestock, Public Water Supply, and Industrial-Manufacturing. In addition to the unique categorical values, there was also an aggregate observation that totaled the water use data across all seven industries. Similarly, there were three water source features that were evaluated with an additional aggregate total feature of all three water sources. The featured water sources included Inland, Great Lakes, and Groundwater. All water use data was reported in number of gallons. Lastly, there were 85 counties that were evaluated for this survey. For the purposes of this model, counties were removed to better understand the specific impact the various industries had on water use. Future studies that include this feature may develop more robust models to forecast future water use.

Michigan population data was also collected from the U.S. Census Bureau from two csv files. The datasets each contained over 14 extraneous feature categories that were summed into a total population estimate. Since population was hypothesized to be an explanatory variable to Michigan’s water use, the total population estimate was the only feature used. Investigating the relationship between different migration categories falls outside the scope of this model. When concatenating the two datasets, 2020 was accounted for twice because it was the transition year between csv files. The population data for 2020 in the newest file (2020 – 2023) was retained as the latest update. The population and water use data sets were merged using year as the common feature.

***Aggregations and reshaping***

One of the main goals for aggregating the data was to distinguish the water source and water use features from each other. After the county and previous index feature were dropped, the DataFrame was melted to create a categorical water source feature and a continuous water use feature. All pre-calculated totals for water source and industries were dropped to ensure data was not overly represented.

Water use and population values were also scaled to represent billions rather than individual units. Since both datasets were estimates of their respective features, reducing these values for higher interpretability was appropriate. After removing counties, the remaining data was also transformed by sum aggregating to represent the total water use for each industry’s water source. The final data transformation was adjusting year to a datetime object for future analysis and modeling. From box plots and value counts, no outliers or null values were identified within each industry. The resulting DataFrame contained 210 observations across five features: year, population (Bil.), water\_source, industry, and water\_use (Bil. gal).

**Data Exploration**

The goal of this project is to forecast Michigan's water use; using industries as a potential explanatory feature. Water use data for each industry will be plotted over the years to understand how it changes over time and how it correlates with population growth. Correlations between industries, water sources, and population must also be assessed to see if there are any underlying relationships between changes over time. Lastly, each feature will be analyzed using the autocorrelation function to show any relationships from the data within itself over time. To better define these goals, three questions were used to guide EDA.

1. How does water use change over time?
2. What relationships exist between industry, water source, and population?
3. Are there annual changes in water usage?

***How does water use change over time?***

To get a baseline for analysis, water use was sum aggregated and compared against the annual population; this was to confirm if there was any underlying relation between increasing populations and water usage.

The Michigan population increased 1.2% from 9.9 million to 10.0 million people (Fig. 1). Compared to the global growth rate in 2022 (0.83%), this is a major increase in the population. In contrast, the water use since 2013 declined by 22.6% (Fig. 1). Although the result contradicts the initial assumption that an increasing population may increase water use demands, a real relationship between population and water use may still exist that can be confirmed using linear regression analysis.

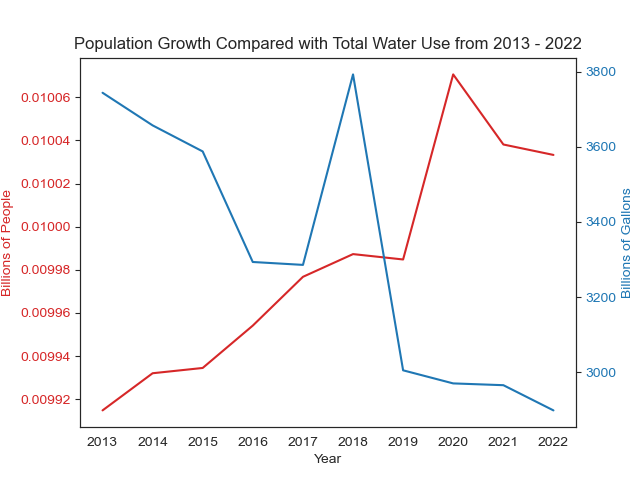


Fig. 1 Temporal distribution of Michigan water use and population estimates.

Looking at the individual industries, only the inland water source for industrial manufacturing and general use; and groundwater for general use increased (Fig. 2). The public water supply from all water sources had the least variability, whereas industrial manufacturing had some of the highest variability over the years. Electrical power generation had the largest water consumption out of all the industries, followed by the public water supply. Both industries were the only ones to be mainly sourced by the Great Lakes. The agricultural industries (irrigation and livestock) pulled most of their water from groundwater and inland sources. This may be due to the vast expanse of low urban areas being further from coastal regions near the Great Lakes.

Despite there being no clear industry contributions (Fig. 2) to the drop in total water use (Fig. 1), there are correlations that will be good to explore. The inverse relationship between population and total water use is of particular interest because it is counter intuitive to human impact on local resources. Additionally, expanding upon the relationship between inland water sources and industrial manufacturing will need correlation tests to confirm the visible trends; similarly, the correlation of groundwater and general use will need to be explored. No annual seasonality between any of the categories have been identified but later autocorrelation will be able to assist with that.

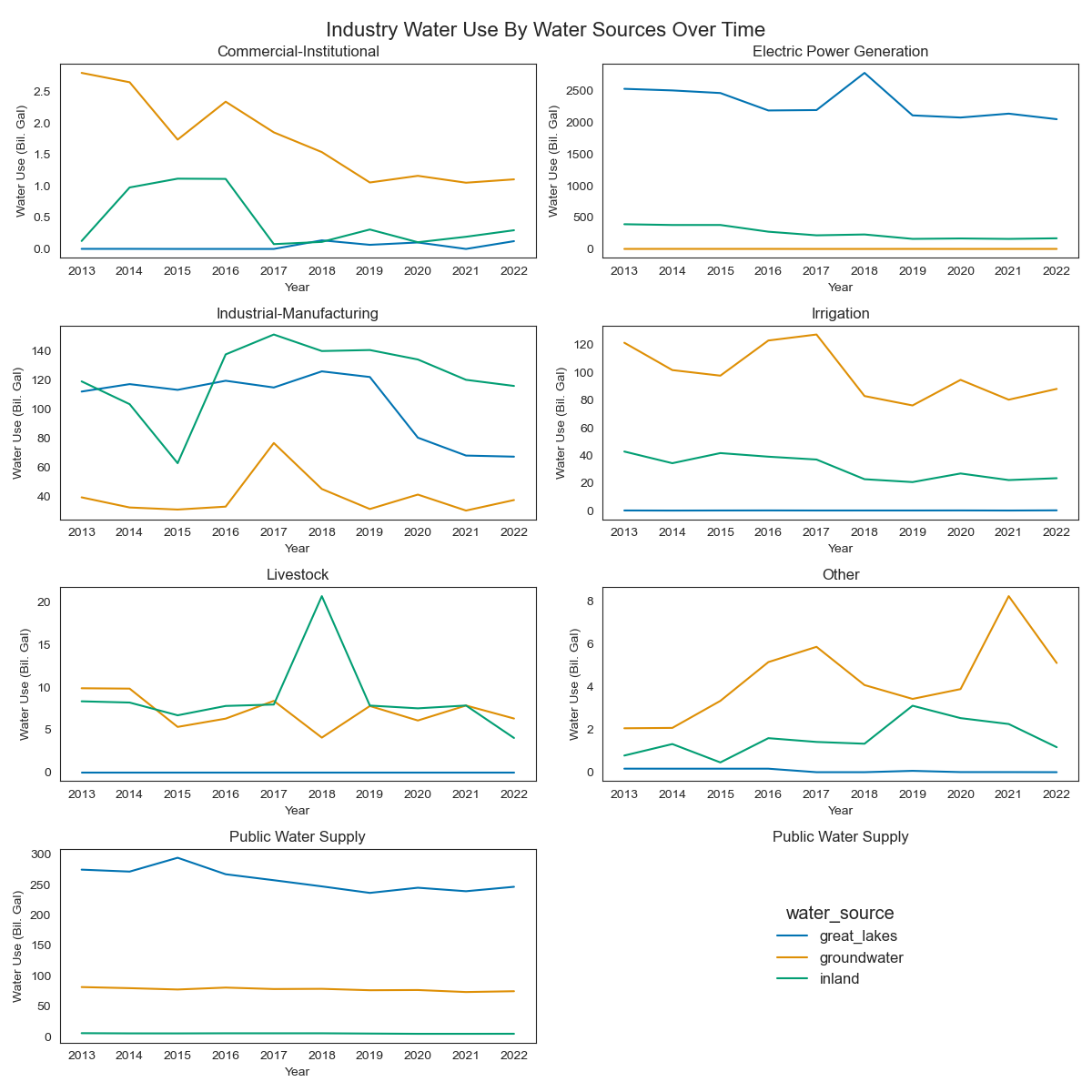


Fig. 2 Temporal distributions of industry water uses distinguished by water sources.

***What relationships exist between industry, water source, and population?***

For this question, five sub questions were identified to segment analysis. Population was assumed to be a key explanatory feature which was supported with the visual linear trends; so, it was compared against each other feature individually. To see if there was any internal relationships between the categorical features, correlation was also tested within each categorical feature.

1. How does population relate to general water use?
2. How does population relate to each industry?
3. How does population relate to each water source?
4. How do the different industries relate internally?
5. How do the different water sources relate internally?

*How does population relate to general water use?*

As expected, there is a significant (p-value = 0.007) negative linear relationship (adjusted r-squared = 0.571) between the population and total water use. The relationship indicates that for every one new person in Michigan, the total number of gallons used drops by about 0.005 billion. The results were further confirmed using Pearsons correlation coefficient (stat = -0.787, p-value = 0.007) suggesting a strong negative relationship between the two features.

The sample size is low, so more archived data and experimentation would be able to reveal any causal relationship. With the total water use relationship established, the different industries should show a similar trend.

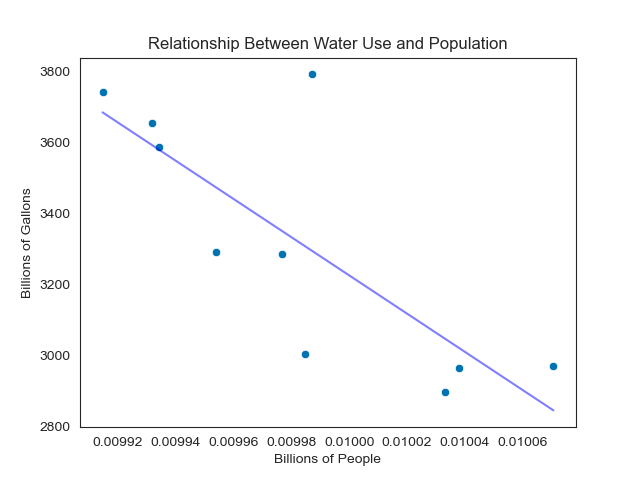


Fig. 3 Linear relationship between water use and population estimates.

*How does population relate to each industry?*

In relation to an increasing population (Table 1), commercial (Pearson = -0.833, p-value = 0.003), public water supplies (Pearson = -0.821, p-value = 0.004), and electric power (Pearson = -0.738, p-value = 0.015) all decrease in water use. General use (Pearson = 0.686, p-value = 0.028) is the only industry that scaled with population growth. The agricultural industries’ variations were not significantly explained by the population growth (Table 1). This may be due to an increased reliance on those industries creating a weak negative linear relationship. Industrial manufacturing also had a weak relationship with population. Industrial manufacturing may account for demand out of state, and so may not be as heavily influenced by the growing population.

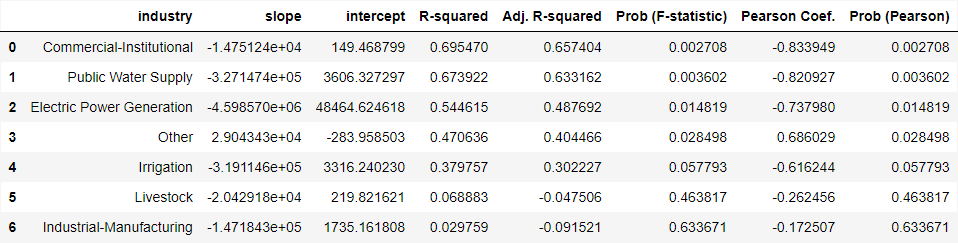


Table 1. Summary correlation statistics between population and each industry.

*How does population relate to each water source?*

Groundwater had the weakest correlation with population growth (Table 2, Pearson = -0.374, p-value = 0.286) which may be represented by the less publicly accessible industries like commercial irrigation. Inland sources showed the strongest relationship (Pearson = -0.907, p-value = 0.000) which may be explained by the geographic distribution of the population. Only so many people can aggregate near the Great Lakes so taking advantage of rivers and lakes across the state will be a good project to investigate further.

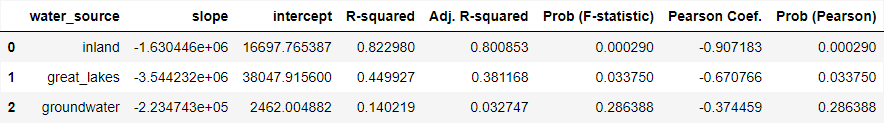


Table 2. Summary correlation statistics between population and each water source.

*How do the different industries relate internally?*

**Modeling**

Four models were constructed to determine which would provide the best classification for the data set. The simple decision tree model (dt) was the first model to train as a baseline comparison to all other models. The dt model is a supervised machine learning (ml) algorithm that can be used as a classifier that selects the tree based on the lowest mean squared error. For this model the tuning hyperparameters were the splitting method ('criterion'), how many layers the trees could have ('max\_depth'), and what the minimum number of samples in each leaf allowed ('min\_samples\_split'). The latter two hyperparameters were included to prevent overfitting.

Next the random forest (rf) model was constructed because it can combine multiple decision trees for a more robust model. Both the dt and rf models were of particular interest because they were able to identify the driving factors of the final results classification. Like the dt model, the number of features present in each tree ('max\_features') was determined to be best set to the square-root of the number of features. The only additional hyperparameter that was tuned was the number of trees that were made ('n\_estimators').

K-nearest neighbors (knn) and logistic regression (lr) models were also trained to more probabilistically classify the final results. Neither of these models provides feature importance identification, but use different methods than the decision trees for classification. The knn model was hypertuned to the number of neighbors (‘n\_neighbors’), the weight distribution (‘weights’), and the algorithm to compute the nearest neighbors (‘algorithm’). The lr model was hypertuned to the penalty calculated to each feature (‘penalty’), the type of algorithm used for optimization (‘solver’), and the regularization strength (‘C’).

***Encoding***

To prepare the data for modeling, the data was encoded since the models cannot process strings efficiently. Student gender (‘gender’) and disability status (‘disability’) were binarily encoded to ‘0’ and ‘1’. Data from this online program did not include any other gender identities so a binary classifier was most appropriate for this feature. Similarly, disability status only checked for whether a student registered as disabled, not the nature of the disability.

The features that were ordinally encoded were education (‘highest\_education’), IMD band (‘imd\_band’), age (‘age\_band’), and final result (‘final\_result’). The IMD and age bands expressed a clear order of progression. For education, the order is the amount of time spent in academia. The final results were also encoded as a progression of rank with 'Distinction' as the top, and 'Withdrawn' in the bottom. Distinction was deemed the highest rank between the four because it is a special honorific beyond passing. Withdrawn was considered the bottom because students who withdrew did not make it to the same standard as the other students to determine their ranking.

The region students were from (‘region’) and assessment activities (‘activity\_type’) were nominally encoded using value counts. To ensure that these features were not underweighted during analysis, they were encoded by their feature value counts. Although the encoded values ranged to the thousands, proportionally, their weights were still representative of the distribution.

***Scaling and Splitting***

Because the data was ordered by assessment originally, the data was shuffled prior to splitting. Additionally, the data was stratified to ensure a proportional representation of the classifications were present in both the training and test sets. Students who passed severely outweighed any other classification. A quarter of the data was reserved as the test set.

The data was scaled using a StandardScaler to ensure all distributions have similar variances for the parametric-based models. The StandardScaler was chosen above a Normalizer or RobustScaler because the outliers had been dealt with during data wrangling.

**Model Results**

Model classifications were compared based on their F1-scores to ensure the most consistent identification of true positives. The F1-score was also chosen because it was accessible by all four classifiers and is more robust to the class imbalance than accuracy.

The best model for classifcation was the K-Nearest Neighbors model (Accuracy = 0.72, Weighted F1-Score = 0.72, Fig 5). Passing students were the easiest to classify for all four models; to the point of mistaking most of the distinguished students for passing classifications. The class imbalance may have severely impacted the prediction of the models to almost default to the passing classification. The knn model is the best model to classify students, however with an accuracy and weighted F1-score of 0.72, there are more tunings required before deploying it as an educational aid for online educational administrators. Being able to predict a student’s trajectory within 72% is very risky. For future iterations of this classification tool, under sampling the passing students to accommodate the class imbalance may improve the model metrics across all models.

Although the knn model provided the best classification, the best decision tree classifier is needed to determine feature importance. Both the dt (Accuracy = 0.62, Weighted F1-Score = 0.60) and rf (Accuracy = 0.64, Weighted F1-Score = 0.59) were very similar when comparing their weighted f1-scores (Fig 5). The rf model was chosen for feature importance over the dt model because it scored higher in accuracy than the dt model.

As anticipated, the average score for each assessment was the strongest predictor (‘mean\_score’ Importance = 0.40) of how a student would end the course (Fig 6). Additionally, the amount of time a student spent on their assignments (‘max\_assessment\_length’ Importance = 0.30, ‘mean\_assessment\_length’ Importance = 0.20, Fig 6) were also strong indicators of how they would be classified by the end of the course. Between these three features, 90% of the classification was predicted. The major distinction between the top three features and the rest may indicate that the program provides equitable enough resources to its students that only their work ethic determines their success (Fig 6).

**Conclusion**

The goal of these models were to create a classification tool that educators could use to predict student trajectories without easy access to them; and to identify the driving factors of how students succeeded or failed. The main determinants identified for this program were how well students scored on average in their assessments and how long they spent completing those assessments. The program indicated through feature importance that it provides equitable resources that enables all demographics to pass or fail by their own work. All demographic features explained less than 10% of the how students would end the course which would be a great indicator of an equitable program.

This project was able to correctly classify, at best, about 72% of the students. The easiest students to classify were passing students who comprised over 50% of the data. To create a more robust model, the class imbalance between how students ended the course needs to be reduced through resampling techniques. The imbalance can be compensated in future studies by subsampling passing student observations or aggregating data across multiple years.

From these findings, administrators of this course may need to invest into maintaining the balance of equitable opportunities for all students and providing sufficient academic support to ensure high scoring. In future models, understanding the features that dictate the higher scores would be crucial to identifying additional resources needed for development. Additionally, more studies are needed to contextualize how assessment duration is associated with students passing or failing the course. In terms of the students who completed the course, they would have had access to more of the scoring credits than students who withdrew so it could be an indication of students simply participating. Other possible explanations are measuring the level of care or dedication to a particular assessment or students not being able to understand the assessments clearly enough. During exploratory data analysis there was an association between the highest level of education and how students finished the courses. Reanalyzing the data controlling for education rather than student or assessment will be a good indicator of how education scales with course placement. Although determining how activity types interacted with the final results became a secondary objective through exploratory data analysis, it was one of the least important features in prediction. Running additional PCA tests would help reweigh the types of activities to better determine if specific activities were more deterministic for students than others.

From this model, driving factors of student success were able to be identified, but the accuracy and consistency of the models would prevent them from being deployed as classification tools. After tuning and resampling, these model results have the potential for equipping online educators to better engage their students.

**References**

Martusiuk, O. (n.d.). Michigan water use data (2013 to 2022) [Data set]. Kaggle. https://www.kaggle.com/datasets/oleksiimartusiuk/michigan-water-use-data-2013-to-2022/data

U.S. Census Bureau. (n.d.). State population totals: 2010-2020 [Data set]. U.S. Census Bureau. https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/state/totals/

U.S. Census Bureau. (n.d.). State population totals: 2020-2023 [Data set]. U.S. Census Bureau. <https://www2.census.gov/programs-surveys/popest/datasets/2020-2023/state/totals/>

U.S. Environmental Protection Agency. (2023). *Climate change indicators: Great Lakes* [Data set]. <https://www.epa.gov/climate-indicators/great-lakes>