JORDAN VALLEY WATER
CONSERVANCY DISTRICT

Water Demand Prediction Model
University of Utah MSBA Capstone Project
Spring 2022 Project

Jackson Roper, Jake Thomas, Nate Hogenson, Zach McDougall

David Eccles School of Business
THE UNIVERSITY OF UTAH

Executive Summary

Jordan Valley Water Conservancy District (JVWCD) is tasked with providing water for much of the Salt Lake Valley. Water demand levels can vary greatly in short periods of time. The sometimes-volatile changes in water demand can often require Operations Staff at JVWCD to work or at least be on standby during overtime hours to service the changing water demands in the valley. There is also a significant cost involved when more water is needed than previously planned for which forces JVWCD to purchase water from other expensive sources to meet the demand.

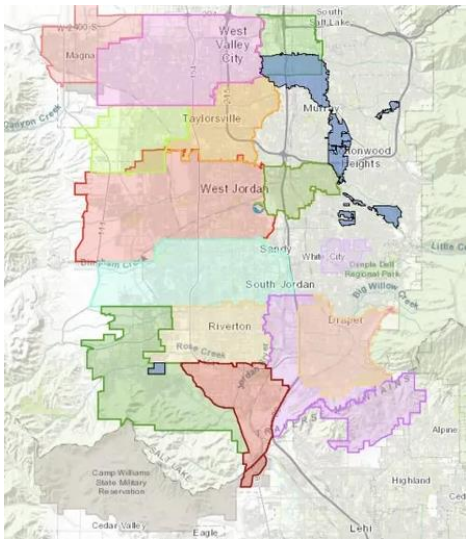
In order to try and mitigate some of these problems, JVWCD wanted to explore water demand prediction models to try and predict future water demands. JVWCD's task for the team was to use weather data collected from the Salt Lake City International Airport to try and find a relationship between the weather variables and water demand in an attempt to predict water demand levels. The main goal of this effort will be to give Operations Staff at Jordan Valley an additional datapoint that will give them more confidence in their daily plans to keep up with the changing levels of water demand. As a combined team, we are hopeful that we will be able to generate accurate predictions that will start to help ease some of the taxing and costly situations that JVWCD faces while servicing water demand in the Salt Lake Valley.

With the data provided by JVWCD, we were able to create the following three models:

- A White Box Linear Regression model with decent predictive power where variable coefficients are readily available.
- A Black Box k Nearest Neighbors (kNN) model with excellent predictive power but no available variable coefficients to explore and analyze.
- A Tableau Time Series Forecasting model that predicts water demand levels by utilizing historic demand actuals to identify seasonal and data trending patterns.

We found that each of these predictive models have their own unique strengths but also have their own unique shortcomings as well. We were able to identify different times throughout the calendar year when each of these models will provide JVWCD with the most value in their daily operations. We are confident that these models will help JVWCD achieve their goals of providing Operations Staff workers with additional data points that will give them more confidence when making important decisions on water usage.

Introduction



Business Problem

JVWCD is tasked with providing water to a large area that includes the majority of the southwestern Salt Lake Valley (See Appendix 1). Their business is interesting as their job is to provide water yet encourage their customers to limit demand for their product as much as possible.

Based on our conversation with the JVWCD team, we understand that they manage a variety of reservoirs throughout the valley that provide water to homes and businesses in their area. Because of municipal safety reasons (such as use by local fire departments), it is imperative that the reservoirs never get too low. In the case that there is too much water in the various reservoirs then they will route it back to the groundwater wells.

The business problem that they would like our help on is that they are finding that too often workers are needing to login on weekends and holidays to make sure they are able to route additional quantities of water into reservoirs due to unforeseen variation in demand. This also involves the cost of last second purchases, but JVWCD made it clear that the main concern is that their team does not have confidence in what the demand will be in the short term. They would like to provide operations staff with a predictive model. This could give their team confidence in their decisions so that it does not need to be watched quite so closely.

With this problem in mind, the JVWCD team provided us with data very early in the project. We were provided with data showing hourly demand and daily weather information taken from the SLC Airport weather station.

The questions our team got to work on answering were these:

- Are weather data variables accurate enough to reliably predict demand?
- Which weather variables best predict demand?
- How can we ensure the model is actionable?

One of the deliverables we had suggested to JVWCD for this project was a recommendation for a good weather data source as we feared that the SLC Airport data would not perform well since it only covers one corner of their area. After researching several solutions, we provided a couple

of options for them and JVWCD informed us that they were confident that the free version of openweathermap.org would be adequate for their needs.

Our initial analysis quickly showed that there were several variables that had strong enough predictive power to create a working model even using the daily SLC Airport data. After we had communicated this information with JVWCD we had aligned for the purpose of this project to simply use the daily SLC Airport as the weather data since it was showing promise for predictability.

In order to find the variables that best predicted demand we went through and made several alterations such as logs, moving averages, and adding additional variables such as seasons to create as much predictive power as possible until we found a model with sufficient predictive power. We found that the predictive power was strong and we moved forward tweaking the model to optimize it as much as possible.

Our model was looking promising. As a result, one of our main concerns is that we would show JVWCD a strong model that they would ultimately stash in a file never to be used again because their operations staff don't have sufficient knowledge and experience in R.

We asked JVWCD how they planned to implement the model since we wanted to make sure they are able to use it. They told us that they already use Tableau and they would like to be able to implement it there. So, we began working on creating a strong model in R so they could implement the coefficients in Tableau for them to use regularly.

This report will detail the results of our exploratory data analysis, search for an accurate model, project deliverables, and reasoning given the context of this project. We believe that the models we have created will provide significant value to JVWCD. Based on our discussion with JVWCD we believe that they have the tools and knowledge to implement our models to be used in daily operations if desired.

Exploratory Data Analysis

JVWCD provided our team with multiple datasets showing the hourly water demand in both cubic feet per second (cfs) and millions of gallons per day (mgd). They also provided data detailing weather information on that day including precipitation, mean/min/max temperature, and snowfall.

Once these datasets were provided, the capstone team performed Exploratory Data Analysis (EDA). EDA would provide the initial assessment on the data and help the team discover characteristics with the data provided. The capstone team gathered a variety of questions during the EDA process that were taken back to the project sponsors for further discussion and input.

The capstone team used the statistical programming language R to analyze the data through a variety of code and queries. Our initial discovery determined several outliers. On three occasions JVWCD experienced negative hourly demand as shown in Figure 1. Does JVWCD really experience negative demand? We found out that demand can be negative or positive if water is returned to the system.

hourly_demand	datetime	datetime_military	date	time	CalendarMonth	CalendarDay	CalendarYear
-49.67661	2021-01-01 12:59:00	01/01/2021 12:59	01/01/2021	12:59	01		2021
-142.6186	2021-03-01 06:59:00	03/01/2021 06:59	03/01/2021	06:59	01		2021
-14.15972	2021-03-05 09:59:00	03/05/2021 09:59	03/05/2021	09:59	05		2021

Figure 1: Negative hourly demand

During the course of the capstone team's EDA, it was determined that a season variable will provide additional insight. Through the use of R, a season column was added to the dataset. As assumed, the summer months have the highest demand values. This is shown in Figure 2.

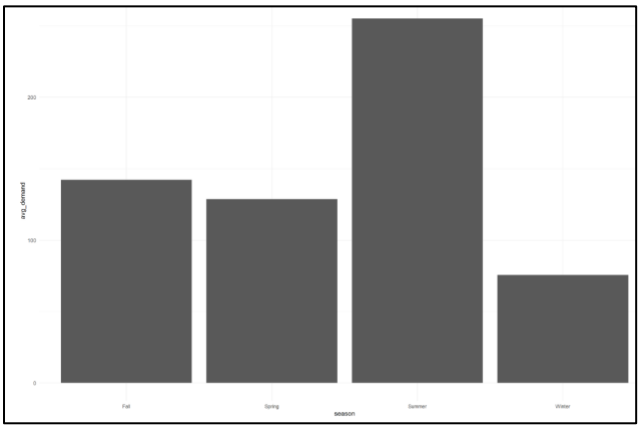


Figure 2: Average demand by season

```
## Rows: 3,072
## Columns: 14
## $ date              <chr> "10/10/2020", "10/10/2020", "10/10/2020", "10/10/202~
## $ TemperatureMax    <dbl> 84.92, 84.92, 84.92, 84.92, 84.92, 84.92, 84.92, 84.~
## $ TemperatureMin    <dbl> 57.02, 57.02, 57.02, 57.02, 57.02, 57.02, 57.02, 57.~
## $ TemperatureMean   <dbl> 70.97, 70.97, 70.97, 70.97, 70.97, 70.97, 70.97, 70.~
## $ Precipitation     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ CoolingDegreeDays <dbl> 5.97, 5.97, 5.97, 5.97, 5.97, 5.97, 5.97, 5.97, 5.97~
## $ SnowDepth         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SnowFall          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ RainDays          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ daily_demand      <dbl> 171.5175, 171.5175, 171.5175, 171.5175, 171.5175, 17~
## $ CalendarMonth     <chr> "10", "10", "10", "10", "10", "10", "10", "10", "10"~
## $ CalendarDay       <chr> "10", "10", "10", "10", "10", "10", "10", "10", "10"~
## $ CalendarYear      <chr> "2020", "2020", "2020", "2020", "2020", "2020", "202~
## $ season            <chr> "Fall", "Fall", "Fall", "Fall", "Fall", "Fall", "Fal~
```

Figure 3: Joined weather data with system demand data

To prepare the data for further analysis, the system demand dataset was joined with the weather dataset collected at the weather monitoring system at the Salt Lake Airport. The resulting dataset had 3,072 rows and 14 columns with the following variables as shown in Figure 3.

Next, the team began looking at a variety of histograms of hourly demand to determine the distribution of the data. Figure 4 shows the hourly demand on the x-axis and how many times that value has occurred on the y-axis.

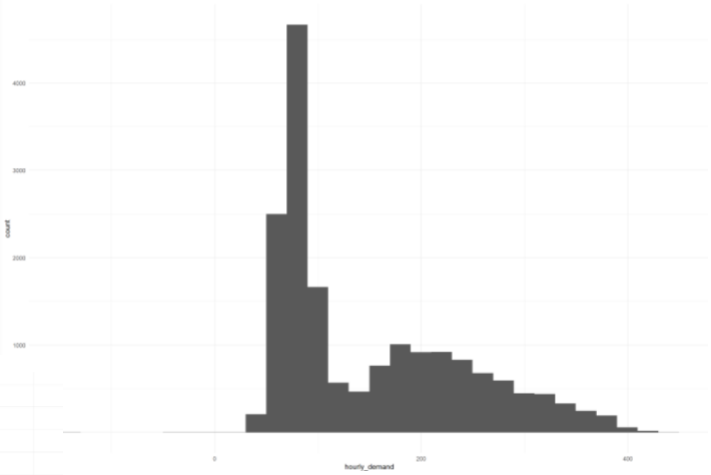


Figure 4: Histogram of hourly demand

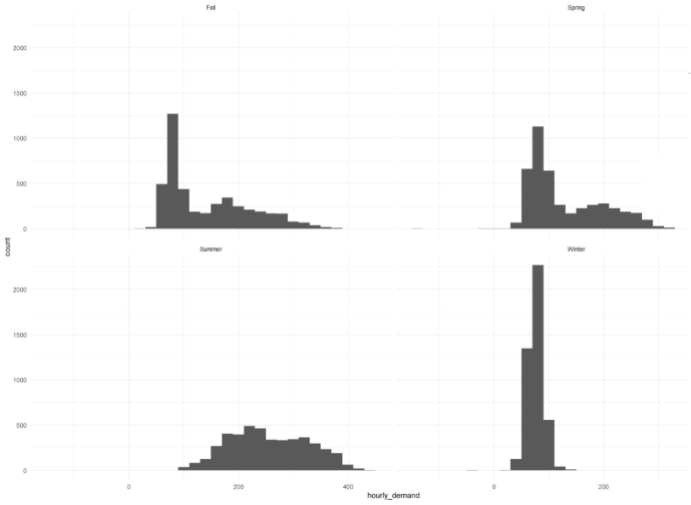


Figure 5: Histograms of seasonal demands

To further see how the data was distributed, hourly demand was split by the newly created season variable. As assumed the summer and fall months saw a higher hourly demand compared with spring and winter.

Thus far in the EDA process we have looked at hourly demand. After discussing the best way of predicting demand, the team along with JVWCD decided that the best view on demand will be in a daily format instead of an hourly. Ultimate goal would be to have a 3 - 5 day forecast to plan for the increase or decrease in water demand. As shown in Figure 6, May through September experienced the highest average daily demand. This coincides with our initial assumption that the temperature variable has an impact on demand.

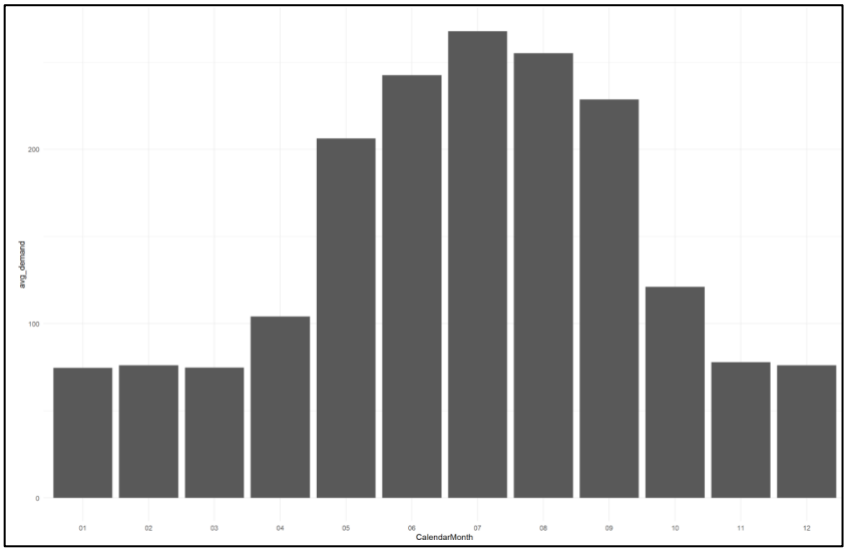


Figure 6: Average demand by month

After the initial data prepping and cleaning was performed, a linear regression model was run with all independent variables. These variables included; date, mean temperature, precipitation, snow fall, rain days, and season. The dependent variable is what we are trying to predict. In this case it is the average daily demand (mgd) for water. Figure 7 shows the performance of the linear regression model. A main indication of performance for a linear regression model is R^2. R^2 measures the proportion of the variance for the dependent variable that the independent variables are able to explain. Essentially, R^2 measures how strong a linear relationship there is between variables. The R^2 for this initial model is 0.8442. This would be considered a strong model. With Figure 7, we are able to see which independent variables are statistically significant. A main measurement of an independent variable being statistically significant or not is the p-value. The benchmark for the industry is if a p-value is ≤ 0.05 it is statistically significant. This would allow you to reject the null hypothesis. The following independent variables are considered statistically significant and provide strong predictive capabilities to predict water demand. These include mean temperature, snow fall, and season. This resulted in more questions than answers for the capstone team. We knew that temperature would be statistically significant, but the team also thought precipitation would be as well. Why did precipitation only have a p-value of 0.090746? The team manipulated the precipitation variable so that instead of looking at the current day precipitation the model would have an independent variable of previous week precipitation. The linear regression model was run with this update. This resulted in previous week precipitation being statistically significant with a p-value of 4.44e-06.

```
Call:
lm(formula = avgdaily_demand ~ ., data = merginated)

Residuals:
     Min       1Q   Median       3Q      Max
-109.705  -21.361    1.291   20.156  105.962

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     651.326164 115.512045   5.639 2.47e-08 ***
ActualDate       -0.038613   0.006154  -6.274 6.07e-10 ***
TemperatureMean   3.842844   0.125073  30.725  < 2e-16 ***
Precipitation   -26.139362  15.432778  -1.694 0.090746 .
SnowFall          4.234368   2.002773   2.114 0.034838 *
RainDays         -1.392541   3.877284  -0.359 0.719586
seasonSpring    -12.611610   3.632204  -3.472 0.000547 ***
seasonSummer     18.969527   4.555853   4.164 3.51e-05 ***
seasonWinter     10.215741   4.494115   2.273 0.023313 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.86 on 717 degrees of freedom
Multiple R-squared:  0.8459, Adjusted R-squared:  0.8442
F-statistic: 492.1 on 8 and 717 DF,  p-value: < 2.2e-16
```

Figure 7: Linear regression model

Next, correlation among the numeric variables was analyzed. Specifically, how strongly or negatively correlated were the independent variables toward daily demand. You want perfect correlation between your independent and dependent variables. It is undesirable to have anything higher than moderate correlation between independent variables. The following correlation metrics could be categorized as follows.

	Correlation
daily_demand	1.0000000
TemperatureMean	0.5495006
TemperatureMin	0.5364903
TemperatureMax	0.5355870
CoolingDegreeDays	0.3853276
SnowDepth	0.0137351
Precipitation	-0.0349699
SnowFall	-0.1267974

Figure 8: Correlation matrix

Perfect Correlation

- If the value we calculate is +/- 1 it is a perfect correlation. This means that as one variable increases if it is a positive the correlated variable also increases. If the variable is negative the correlated variable decreases.

Strong Correlation

- If the variable has a value between +/- .50 to +/- 1 there is a strong correlation among the variables.

Moderate Correlation

- If the variable has a value between +/- .30 to +/- .49 there is a medium correlation among the variables.

Low Correlation

- If the variable has a value below .29 there is small correlation among the variables.

No Correlation

- If the variable has a value of 0 there is no correlation among the variables.

Prior to finalizing any predictive model, we must test the dataset for multicollinearity. Multicollinearity occurs when two or more independent variables that are used for prediction are associated with another. If left alone, multicollinearity can cause inaccurate readings when examining your coefficients and reduce the influence of your model to identify predictor variables that are statistically significant. It could also cause your model to be overfitted.

We will focus on analyzing and removing independent variables that have a strong correlation with one another.

To better visualize the correlation among all numeric variables, we use the function corrplot(). This plots the data and clearly shows which variables have a strong correlation and which do not. Additionally, the plot allows us to see which variables are positively correlated (blue) or negatively correlated (red).There is a strong correlation between the max temperature, mean temperature and minimum temperature. This will be addressed to prevent the model from overfitting. We will remove these variables later on so that we have accurate relationships between all independent variables.

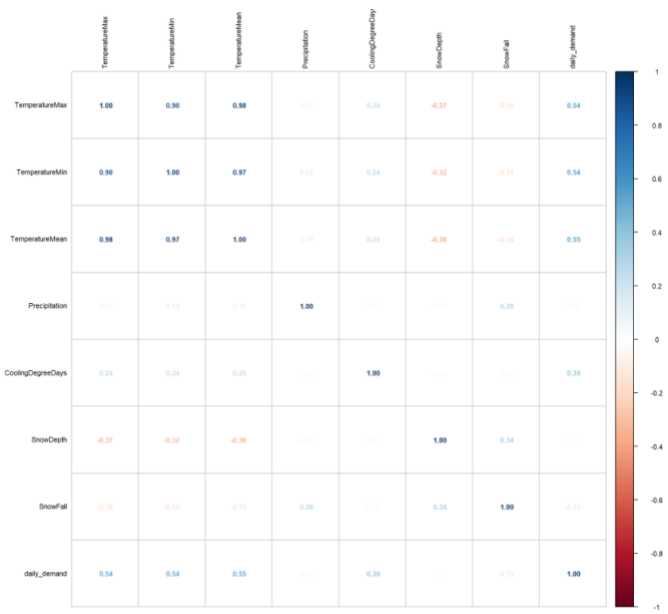


Figure 9: Correlation plot

The Search for an Accurate Model

Now that we had an understanding of our data, we wanted to see which variables had the highest predictive capabilities. It made logical sense that most of the weather variables such as temperature, precipitation, and season should provide value to the prediction model. We made sure to add the variable "Season" by grouping dates into factors of summer, spring, winter, and fall. We created simple regression models using the original data as well as altered data and found that some variables had very high p-values.

As mentioned earlier, we were particularly confused about precipitation not providing a low p-value since one would think that more natural water would result in less need for water from reservoirs. Logically we decided that we use less water when it has been a rainy week, but not necessarily because it rained that day. So, we went and created a new variable that took the average precipitation for the previous week and found that it had a p-value far below 0.01. In our discussion with the JVWCD team they suggested that we go with something more like an average of the last 2-3 days instead of the previous week. With this in mind we decided to test it out. We took the R^2 of the model using the average precipitation from the previous 1 day all the way to the previous 14 days and took the R^2 of the train and test data set. From this we found that the highest test R^2 we could get before we start to see it dip back down is about 9-10 days (see visual above). Based on that information we decided to use 9 days for the moving average.

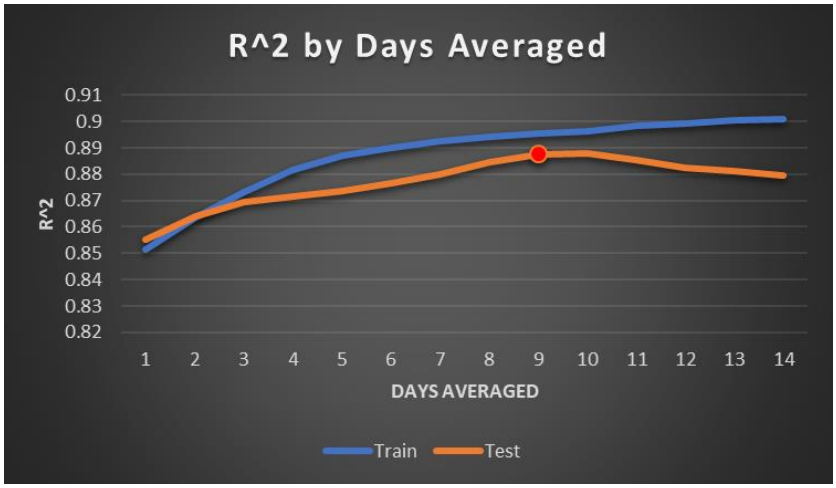


Figure 10: R^2 for linear model taking averages of previous n days

Test

With this model created we were hopeful for the predictive ability of our model. As you can see in the above visual, we were seeing an R^2 over 0.88 in both the train and test sets that we had split and randomized. Eager to see the model perform with the presentation only weeks away, we reached out to JVWCD and asked them to provide us with their latest demand for this year so we could use our model to show how well it predicts. However, to the shock of the entire team the model performed very poorly. The R^2 value for the first part of 2022 was 0.14 which is significantly lower than what we had received on our test dataset.

```
[1] "2022"
        R2        MAE       MAPE       RMSE
0.1418951 22.7817602 44.5043856 26.3033051
```

Figure 11: Performance of 2022 predictions

We quickly got to work searching for what the problem might be. Some of us were convinced it had to be some sort of mistake or that we weren't interpreting the results correctly (no way we were the problem, right?). After looking through the results, what we ultimately found was that although our model was showing strong results overall, it was struggling to make accurate predictions in the winter months. We broke up the model to see its performance in each season (see Figure 12) and found it to perform well in the spring and fall, acceptably in the summer, and abysmally in the winter. Having discovered this, our team decided that although the model generates decent overall predictions, the model really should work for the entire year to be actionable in a way that is meaningful and adds value to JVWCD.

```
[1] "Spring"
        R2        MAE       MAPE       RMSE
 0.7442329 16.3587889 24.5346923 19.8060826
[1] "_____________________________________________"
[1] "Summer"
        R2        MAE       MAPE       RMSE
 0.6164364 13.8317237  9.1803469 16.6123394
[1] "_____________________________________________"
[1] "Fall"
        R2        MAE       MAPE       RMSE
 0.8988446 11.4349962 15.7674933 14.5619337
[1] "_____________________________________________"
[1] "Winter"
         R2         MAE        MAPE        RMSE
 0.00417281 12.36149540 25.06113806 17.14750888
```

Figure 12: Performance of seasonal predictions

<u>Looking Beyond Linear Regression</u>

Seeing the less than stellar 2022 and wintertime performance from our linear regression model, we decided to try out additional algorithms to see if we could improve upon our modeling performance. Using the same input data for our linear regression model, we created multiple models using multilayer perceptron (MLP), k Nearest Neighbors (kNN) and Support Vector Machine (kSVM) algorithms. We saw similar performance results in the MLP and SVM algorithms and to our surprise, much improved performance results in the kNN algorithm (See Figure 13). The new challenge that the team faced was to try and understand why the kNN model generated better predictions compared to other models that we tested out. The linear regression model demonstrated fine performance outside of the wintertime. What made the kNN model perform well year-round?

```
[1] "Train"
      R2       MAE      MAPE      RMSE
0.9951177 2.8676376 3.5063882 4.0287700
[1] "_____________________________________________"
[1] "Test"
       R2        MAE       MAPE       RMSE
 0.9103319 11.2717474 14.1593598 15.9359505
[1] "_____________________________________________"
[1] "2022"
     R2      MAE     MAPE     RMSE
0.668055 1.615680 3.104153 2.559372
[1] "_____________________________________________"
[1] "Spring"
      R2       MAE      MAPE      RMSE
0.9575133 6.1102763 8.1319615 9.4584472
[1] "_____________________________________________"
[1] "Summer"
      R2       MAE      MAPE      RMSE
0.9256647 4.7332502 3.3379667 8.0723136
[1] "_____________________________________________"
[1] "Fall"
       R2        MAE       MAPE       RMSE
 0.9625551  6.7544996  8.8261196 10.5075767
[1] "_____________________________________________"
[1] "Winter"
      R2       MAE      MAPE      RMSE
0.8420095 1.3443753 2.7412899 2.0279301
```

Figure 13: kNN model performance

kNN Explanation

Our theory on why the kNN model performs better than the other models we tested has to do with the characteristics of kNN modeling. In a nutshell, linear regression tries to find a line that best fits the data points in a given dataset. The main downfall of trying to predict water demand using linear regression is that water demand values are not linear. kNN modeling instead sorts similar data points (neighbors) into neighborhoods and then uses those neighborhoods to make predictions on new datasets (See Appendix 2). The analyst training the kNN model is able to select (or use the algorithm to optimize) the size of the neighborhood (how many neighbors make up a neighborhood). In our case, the best kNN model performance came when we selected 34 as the size of the neighborhood (k = 34). Because weather doesn't play as much of factor during the cold seasons, linear regression models struggle with performance because they rely upon the weather variables to make predictions. kNN modeling should help to fill that void because it uses similar data points to make predictions.

JVWCD has found that, generally speaking, the public in the Salt Lake Valley is responding positively to water conservation messaging that the local government has published. Because of historic drought conditions that the state is currently facing, there is a chance that 2022 water demand levels in Utah will vary greatly compared to previous years. There is also the chance that weather patterns will not play as big a factor in demand levels due to the general public making concerted efforts to use less water. kNN modeling may then have to be relied upon more to make accurate predictions during these times of uncertainty.

Tableau Time Series Forecasting

Because JVWCD utilizes Tableau in their daily operations, we also wanted to explore water demand predictions using the "Forecast" functionality in the Analytics pane that Tableau offers to users. "Forecast" differs from the previously discussed regression analyses in the fact that Tableau relies upon built-in Time Series Forecasting algorithms to make predictions rather than the predictor variables (temperature, time of year, precipitation, etc.) that regression and kNN use to make demand predictions. Time Series Forecasting algorithms focus on Trend (the tendency in data to increase or decrease over time) and Seasonality (repeatable, predictable variation in data points [ex: water demand levels throughout the year, seasonal temperatures, etc.]) to make predictions. Generally speaking, the more historical data points that the Time Series Forecast contains, the better off the predictions will be.

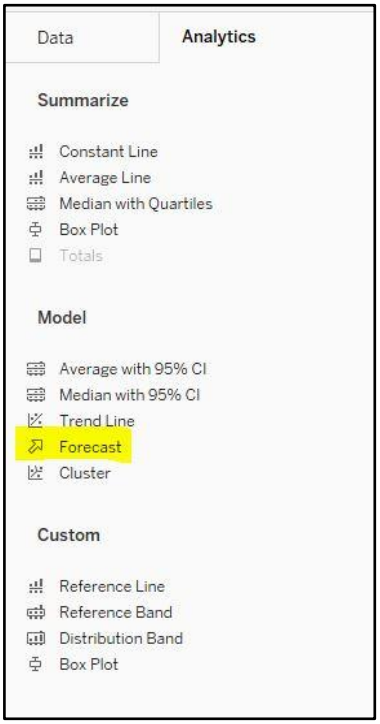


Figure 14: Tableau forecasting

During the process of generating Time Series predictions, Tableau tests 8 built-in algorithms and selects the algorithm that provides the highest quality forecast (See Appendix 4 for additional details). What Tableau Time Series Forecasting lacks in sophistication and complexity, it makes up for with simplicity, and overall easy of use/implementation. The demand forecast below only took a couple of minutes to generate.

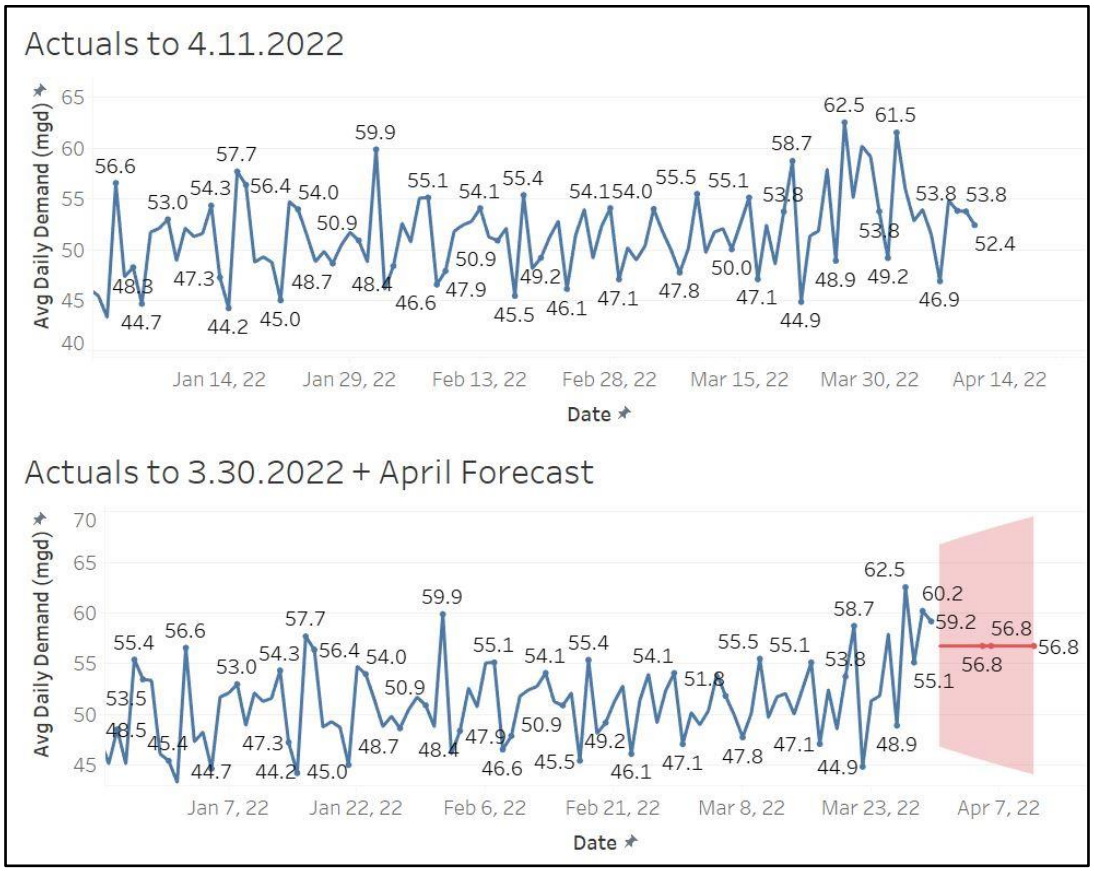


Figure 15: Actuals vs forecast

In the above visualization, the top graph shows average demand (mgd) actuals through April 11, 2022 and the bottom graph shows average demand (mgd) actuals through March 30, 2022 plus the forecasted average demands for April 1 - April 11, 2022. The pink area surrounding the forecast shows a 95% Confidence Interval of where possible average demand values could be. To further illustrate how effective this forecasting method can be, let's take a look at the actual vs predicted values.

Date	Actual (mgd)	Prediction
4/1/22	49.19	56.76
4/2/22	61.54	56.76
4/3/22	56.00	56.76
4/4/22	52.85	56.76
4/5/22	53.96	56.76
4/6/22	51.49	56.76
4/7/22	46.91	56.76
4/8/22	54.83	56.76
4/9/22	53.84	56.76
4/10/22	53.77	56.76
4/11/22	52.42	56.76

Figure 16: Tableau predictions in detail

Something to keep in mind with the Tableau Time Series Forecast is that (like weather forecasts), short term forecasts will generally be of more use than long term forecasts because the forecasts are generated from the near-term data points. Tableau will provide a short description of the data points used to generate the forecast. As we can see in Figure 17, Tableau used data from December 1, 2021 to March 30, 2022 to generate the forecasted demand values for April 2022.

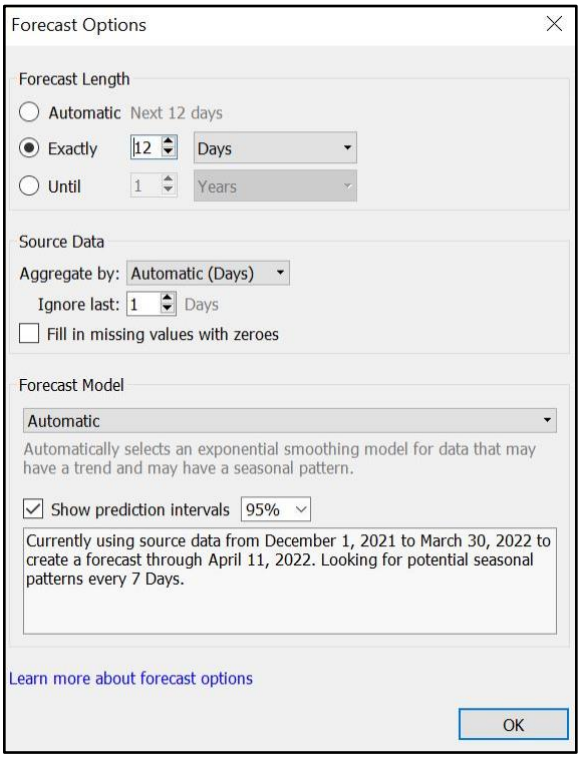


Figure 17: Tableau forecast options

As mentioned previously, implementing Tableau's forecasting functionality into JVWCD's daily operations will likely be a lot easier to accomplish compared to the aforementioned regression predictions. After generating a time series plot(s) with forecasted demand values, the next step in the implementation process is to connect the Tableau dashboard to a live data source that is updated periodically. This should then give JVWCD a live forecast that updates as new data is gathered. While simplistic in nature, these forecasts will be an additional valuable data point that Operations Stuff at JVWCD can rely upon in their daily operations.

Conclusion

Circling back to the questions posed in the introduction we believe we have answered the first two comfortably.

1) Given the right model the variables in these data provide sufficient predictive capability to create an accurate model.

From the data we were provided we were able to create a model with significant promise that shows predictive capability on not only the test data, but also this year's data given to us by JVWCD.

2) The variables that seem to predict demand the best with the most significant impact are the season and the precipitation. Precipitation is most effective at prediction when using a rolling 9-day average. However, multiple variables from max temperature to days of rain in a week provide valuable insight for demand prediction.

There is clearly a strong correlation and a strong predictive ability from variables in the weather data for predicting future demand.

3) How can we ensure the model is actionable?

We are confident that our model is capable of predicting and we are confident that the skills of the JVWCD team are sufficient to implement our prediction models, but the real-world efficacy of our models have yet to be proven.

Discussion

<u>Benefit to JVWCD?</u>

Objectively, the models are showing promising performance as well as providing insights on coefficients that predict demand. JVWCD mentioned multiple times that their goal with this project is for us to provide a model that gives their workers confidence in making decisions for demand. We expect that the models will perform well in that regard. We are also confident that the coefficients extracted from the linear regression model can give the JVWCD team valuable insight on which variables have predictive capability or how they can be altered to provide predictive capability.

Future Data

A question we still have for future analysis is how usual or unusual were the years 2020 and 2021? The historical demand that JVWCD provided for this project began at Jan-2020 and ended mid-April 2022. We would like to see how much weather varies over time and what effect that might have on the predictive capability of a model. From there it would be interesting to see if we could possibly create a precipitation by season variable or even by year to capture variation that might occur from droughts. We would also like to see if this model could improve when given more years of information to work with so that it captures less noise and better predictive coefficients.

Predictive Capability of New Weather Data Source openweathermap.org

We would also like to see if we could improve the model using more accurate data or create multiple models for different regions in the valley to give more specific predictions. It seems logical that the statistically significant variables might not change too much, but would like to see if we could reduce model noise by breaking up into regions.

Appendix

1. Image taken from Website: https://www.fox13now.com/news/local-news/jordan-valley-water-conservancy-district-proposes-180-million-property-tax-hike-for-infrastructure
2. kNN Explanation: https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3
    a. Disclaimer: The code examples in this article are given in Python. The article still does a really job of explaining kNN modeling in simple terms and you will see kNN modeling in our master code file included in the deliverables.
3. Tableau Time Series Forecasting Details: https://help.tableau.com/current/pro/desktop/en-us/forecast_how_it_works.htm

Definitions & Brief Explanations

R^2: Tells us that the predictor variables in the model are able to explain x% of the variation in target variable.

MAE: Is to subtract our predicted value and actual value at each time point to obtain the absolute value, and then average it out.

RMSE: Tells us the average deviation between the predicted demand made by the model and the actual demand value.

MAPE: returns error as a percentage, making it easy to understand:

MAPE Interpretation

< 10% = Very Good
10% - 20% = Good
20% - 50% = OK
> 50% = Not Goo

Modeling Performance

Metrics:

Linear

	R2	MAE	MAPE	RMSE
Train	0.905363	12.98234	17.67125	16.20255
2022	0.148067	14.57398	28.17016	16.71489
Spring	0.875949	16.0807	23.42733	19.14584
Summer	0.752083	15.27647	10.10531	18.52248
Fall	0.952126	11.68395	15.67997	14.79749
Winter	0.02659	10.93615	22.53676	14.53596

Multi-layer Perceptron

	R2	MAE	MAPE	RMSE
Train	0.975044	8.317984	11.78958	10.59732
2022	0.188556	5.916363	11.73664	7.370434
Spring	0.952968	9.418534	14.56062	11.84
Summer	0.813123	10.05181	6.838171	13.04797
Fall	0.959627	8.494776	11.58868	10.73025
Winter	0.171542	6.442128	13.71555	7.511728

K-Nearest Neighbors

	R2	MAE	MAPE	RMSE
Train	0.996104	2.588474	3.155685	3.594346
2022	0.676855	1.5761	3.019523	2.560977
Spring	0.964084	5.550111	7.466642	8.736907
Summer	0.938612	4.428986	3.121006	7.464391
Fall	0.967966	6.119656	7.962533	9.704927
Winter	0.836695	1.318637	2.69129	2.017056

ksvm: Support Vector Machines

	R2	MAE	MAPE	RMSE
Train	0.988328	4.572242	6.064881	5.694341
2022	0.101532	3.376048	6.591816	3.906666
Spring	0.968282	5.428191	8.131492	6.910414
Summer	0.917182	5.997301	3.935264	7.721319
Fall	0.978608	5.189912	6.490383	6.646012
Winter	0.237395	3.554509	7.29304	4.580294