Kevin Ash, Jacob McKinney, Parker Banks

Dr. Quigley

CSCI 4622: Machine Learning

6 May, 2023

Predicting I-70 Traffic using Weather Data

Problem Space

The goal of our project is to develop a machine learning model that can accurately predict the amount of traffic passing by a measuring station along I70 for a given hour of the day. To achieve this, we have utilized data from the National Oceanic and Atmospheric Administration (NOAA) for daily snow totals and snow depth from a measuring station in Winter Park, Colorado, as well as data from the Colorado Department of Transportation (CDOT) with a counter of the number of cars passing by the station (count station 000120 on I70 right before Idaho Springs).

Our solution aims to address the need for accurate traffic predictions, which can be valuable for transportation planning, traffic management, and infrastructure optimization. By considering the impact of snowfall and snow depth on traffic volume, we aim to enhance the understanding of traffic patterns and improve traffic flow management in the I-70 area.

Our approach differs from existing methods in a couple important ways. First, our predictions are made by the hour, whereas most previous research has predictions of traffic volume by the day[1]. Second, we made the prediction at a particular station along I-70. This prediction could be done for any of the many stations. Previous work only displays data about I-70 as a whole, which makes our predictions much more specific.

Approach

We followed a multi-step approach to address the problem of traffic prediction along I70. The key steps in our approach are as follows:

Data Collection: We collected data from multiple sources. We used NOAA for snow-related data, and CDOT for traffic volume data. We had to dig through a large amount of information that was not relevant.

Data Transformation: We processed and transformed the raw data to create a formatted data frame that combines the relevant information. This involved handling missing or invalid data, filtering traffic data based on the direction of interest, summing up daily traffic count, and extracting relevant features such as month, year, day, day of week, hour, snow depth, and daily snowfall.

Data Splitting and Scaling: We split the transformed dataframe into training and testing sets using the train_test_split function from the scikit-learn library. This division allows us to train the model on a subset of the data and evaluate its performance on unseen data.Then, we standardized the numerical features by applying z-score normalization using the StandardScaler from scikit-learn. This step ensures that all features have a similar scale and prevents certain features from dominating the learning process.

Model Selection and Training: We chose to employ a neural network model using the TensorFlow library. Specifically, we built a sequential model using ReLU activation with multiple dense layers, incorporating dropout regularization to prevent overfitting. The model was trained using the Mean Squared Logarithmic Error (MSLE) loss function and optimized using the Adam optimizer.

Model Evaluation: We trained the model for 20 epochs, with a batch size of 64. During the training process, we monitored the loss and MSLE metrics on both the training and validation sets. We evaluated the model's performance based on its ability to accurately predict the traffic count within a range of tolerances of the actual count.

Data

The data used for our project includes snow-related information from NOAA and traffic volume data from CDOT. For each year (2019-2022), we collected the following datasets:

Snow depth data (Winter Park, Colorado)[2]: This dataset contains the daily snow depth measurements recorded at a specific measuring station in Winter Park. An example is shown here:

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	23	32	37	40	12	0	0	0	0	0	10	18
2	22	32	42	40	11	0	0	0	0	0	9	17
3	22	31	50	40	11	0	0	0	0	0	9	17

Daily snow data (Winter Park, Colorado)[2]: This dataset provides the daily snowfall totals for Winter Park. An example is shown here:

Day	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0	0	T	0	3	0	0	0	0	0	0	1
2	0	0	4.9	0.6	T	0	0	0	0	0	0	0
3	0	0	11.5	2	0	0	0	0	0	0	0	0

Traffic volume data (I70, Idaho Springs)[3]: This dataset includes the hourly traffic count measurements recorded at the count station 000120 on I70 right before Idaho Springs. An example is shown here:

1	COUNTSTATIONID	COUNTDATE	COUNTDIR	HOUR0	HOUR1	HOUR2	HOUR3	HOUR4	HOUR5	HOUR6	HOUR7	HOUR8	HOUR9	HOUR10
2	000120	20190101	P	85	86	93	101	135	214	354	658	1006	1578	2532
3	000120	20190101	S	127	144	105	73	123	207	953	2249	1969	1850	2122
4	000120	20190102	P	152	96	109	115	190	421	854	1234	1398	1692	2180

Results

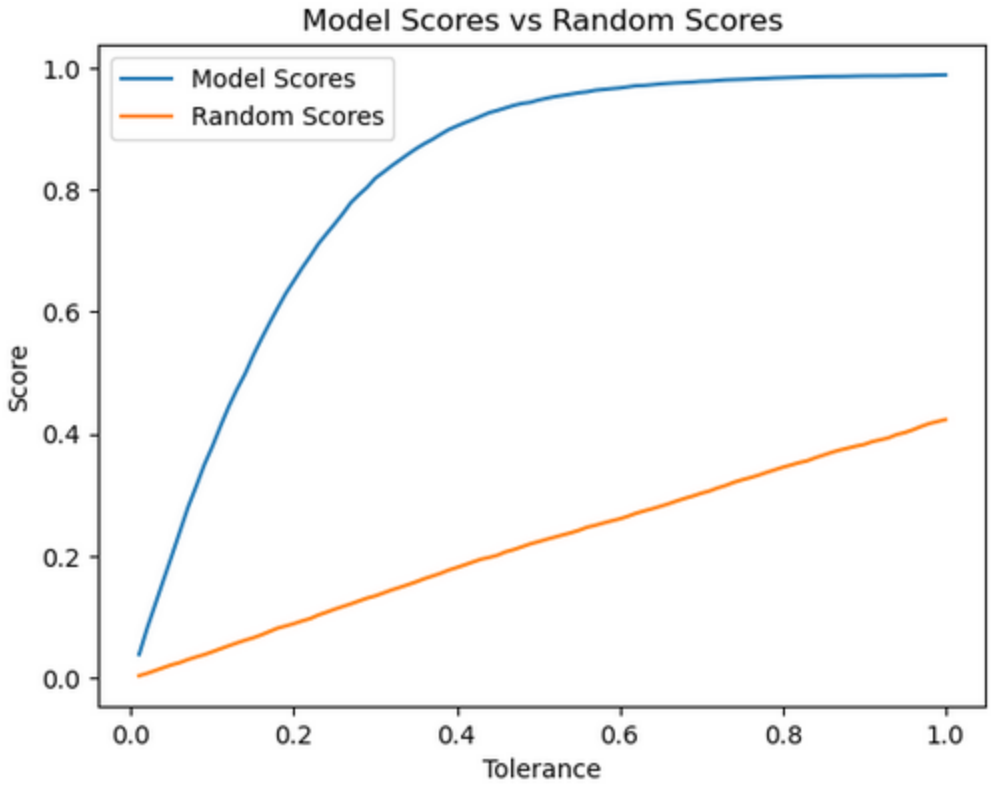
We found that our regression model performed well at predicting the hourly traffic volume. We measured this by finding its accuracy given certain tolerances, as well as calculating the R-squared value, Adjusted R-squared value, and Root Mean Squared Logarithmic Error.

First, we looked at the accuracy of our predictions within certain tolerances. This worked as follows: given a y-value from the test set, we created a range of acceptable values using a tolerance. A tolerance of 0.2, for example would accept values within 20% of the y-value in either direction. We also created a set of random values for comparison. The random values would be between the minimum and maximum values for the true labels in the test set. This way, we could look at the comparison of our model to random guesses of what the traffic count would be. Note that a tolerance of 100% does not equate to a score of 1.0 for the random values, because the tolerance represents 100% either direction of the actual value, which does not necessarily cover the entire range of possible values.

At a tolerance of 10%, the random score was 0.0585, indicating that the randomly generated values fell within the actual traffic counts only 5.85% of the time. In contrast, the model score at the same tolerance was 0.3801, demonstrating that the machine learning model

achieved a significantly higher accuracy of 38.01%. For a tolerance of 20%, we had a corresponding random score of 0.1153 and model score of 0.6678. And, finally, for a tolerance of 40%, the random score was 0.1293, whereas the model score was 0.9064. This means that our model could guess within 40% of the correct value 90% of the time, which is a high level of accuracy.

We also looked at the R-squared value and Adjusted R-squared value. The R-squared value for our model was 0.8758, whereas the Adjusted R-squared value was 0.8757. From the R-squared value, we can tell that 87.5% of the variability in the dependent variable can be explained by the independent variables in the model. The Adjusted R-squared value suggests that the model has a good fit and is able to explain a substantial amount of the variance while accounting for the complexity of the model. It was also barely lower than the R-squared value, which tells us that the predictors included in the model are relevant and contribute meaningfully to explaining the variability in the dependent variable. This also suggests that the model is not overly complex or overfitting the data.



Finally, we looked at the Root Mean Squared Logarithmic Error, which had a value of 0.33641. This means that the average squared logarithmic difference between the actual and predicted values is fairly low. We chose to use this because we had a large range of strictly positive values.

Discussion

The results of our study indicate a fairly strong relationship between our independent variables and traffic volume. Some of the most important independent variables are day of week, hour, and daily snow total. The model still performs fairly well in the absence of the other dependent variables. This makes sense; day of the week is a strong predictor of traffic, as is the hour of day. Snow total also has a significant impact, as more snow on the roads has the potential to significantly lower the amount of cars passing a station in an hour.

These results could have a fairly significant impact on the field. Given a few easy to obtain variables, it allows accurate prediction of traffic along roadways. Furthermore, it does so by station and hour, which could be used to create accurate and specific traffic maps for the future. This could be useful for everyday people for planning for drives, but also has important implications for infrastructure planning and management.

In the future, we would like to explore the effect of other variables such as construction and public events. This data would be harder to find, but would likely greatly impact the traffic flow. Ultimately, we could use this information to create an application for consumers that would allow them to view predicted future traffic.

Link to Repository:

https://github.com/ashkevin22/I70TrafficPredictor

Works Cited

(1) “Travel Forecast.” *goi70*, CSBOX, 4 May 2023, https://goi70.com/travel.

(2) “Daily Summaries Station Details.” *Daily Summaries Station Details: WINTER PARK, CO US, GHCND:USC00059175 | Climate Data Online (CDO) | National Climatic Data Center (NCDC)*, National Oceanic and Atmospheric Administration, https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USC00059175/detail.

(3) “CDOT-Otis Online Transportation Information System.” *Traffic Data Explorer*, Online Transportation Information System, https://dtdapps.coloradodot.info/otis/trafficdata#ui/2/0/0/station/000120/criteria/000120/.