# Prediction of Average Annual Daily Traffic Using Machine Learning Methods

Srinivasan Suresh[#1]

*#1 Student, Doctorate in Computer Science, Aspen University, Denver, Colorado, USA*

***Abstract:***

*The field of machine learning is growing enormously and been widely used. There are many techniques been developed in the machine learning arena. Using these techniques to predict the Average Annual Daily Traffic (AADT) would help to improve the accuracy in prediction and to plan routes in a better manner. The Linear regression, Ridge regression and Lasso regression methods are used in this analysis. The data is obtained from the Highway Performance Management System (HPMS) and from Virginia Roads, the official data provider for the State of Virginia. The raw data is cleansed, regrouped and prepared for the regression analysis. Route category, Through lanes, Facility type and Functional class are the key variables used in the analysis. The machine learning models built through these methods have almost 77% accuracy and these models can be reused to predict the AADT values for new routes or extension of routes.*

**Keywords —** *Annual Average Daily Traffic (AADT), Machine Learning, Through Lanes, Functional System*

## I. INTRODUCTION

AADT is the average volume of traffic across all days for a year on a section of the route. The traffic on a route is calculated by many methods. In simple average method, the traffic volume is obtained by placing an electronic counter and sensor wire. As vehicles pass over the sensor wire, it is reflected in the counter and accumulated in the count. The AASHTO method uses 84 averages, which comes from 7 averages from days of the week, for 12 months (U. S. Department of Transportation, 2018). The cost involved in collecting the traffic data for an entire year is a great expense to the agency. The workforce involved in measuring the traffic, the equipment and maintenance adds costs to it. Usually in cities, the traffic volume is high, as the population is denser and people would commute more. In villages or rural areas, the traffic is expected to be low until Interstates or primary routes pass through. The Virginia state maintained routes are recordedalong with its attributes for various utilities. The total mileage of the routes maintained by VDOT is 93,127 Kilometers. Along with the location of all the routes, the information related to the routes like Number of lane, the surface width of the route, Median etc. is also gathered and maintained.

## II. LITERATURESURVEY

The AADT prediction is been performed for many years. Sliupas, T. (2006) has performed the AADT prediction and forecasting using simple regression and multiple regression methods for the State of Idaho, USA. The study is done for the Lithuanian highways. Traditionally, the forecasting is done with a growth factor using a mathematical formula. The prediction through the regression methods and Idaho method were done and compared with the growth factor method. The predictions were made for an Average scenario, Optimistic scenario and Pessimistic scenario. The forecasts from multiple resources were taken and compared with the results. The predictions from Idaho method were closer to the obtained values.

Selby, B. and Kockelman, M. K. (2011) performed spatial prediction of the AADT value using the ArcGIS software. Speed limits, Functional class, Facility type, Number of lanes, urban area type are some of the parameters used in the analysis. The transportation data from the State of Texas is used in this analysis. For the spatial interpolation both the Euclidean distances and Network distances were used. In few cases where the traffic count is low or the population is less, some errors were observed. The universal kriging method provided more accurate results compared to the spatial regression methods.

Wu, J. and Hao, X (2019) have worked on models to predict the AADT value for minor roads at intersections. The data is obtained from the State of North Carolina. Urban type, Functional class, Number of lanes and Population are used in the analysis. Random Forest method and Multiple Linear Regression methods are used in their analysis. The estimation is made for different type of routes including rural two lane, rural multi lane highway, urban roads and streets. The R software package is used for the regression analysis. After comparing the accuracy from both the methods, the Linear Regression method was recommended for the analysis.

Chowdhury et al. (2019) have developed cost effective methods to predict the AADT value. Least square regression, Factor method and Support Vector regression methods were used in this process. The data is gathered from the 164 permanent count stations installed on Highways, Arterials and Locals. The models are built in Python with Selenium library.

## III. DATA PREPARATION

### *A. Data Source*

The Federal Highway Administration (FHWA) agency supports the State agencies through financial and technical aids. The Highway Performance Management System (HPMS)is managed by the Federal Highway Administration agency and requires the State agencies to collect and submit the data on all of the routes maintained by the Agency. This data is available to the public and it is used in this research analysis(Virginia Department of Transportation, 2019).

The Department of Transportation in the State of Virginia collects the data of all the routes maintained by the agency and various features of the road. This data is reported annually to FHWA.

### *B. Data description*

The explanatory variables considered for the analysis are the Route category code which explains the type of route, Functional system, Facility type, Number of lanes, Toll type and Toll charged direction. The different values of functional system and the count of data are shown in Table 1.

**TABLE 1**
**Summary of Functional system values in the dataset**

| Functional System Number | Functional System Description | Count |
|---|---|---|
| 1 | Interstate | 16818 |
| 2 | Principal Arterial – Other Freeways and Expressways | 12780 |
| 3 | Principal Arterial - Other | 116604 |
| 4 | Minor Arterial | 233772 |
| 5 | Major Collector | 259025 |
| 6 | Minor Collector | 90 |
| 7 | Local | 518 |

### *C. Data preprocessing*

The Route Category code, which is a categorical variable, is grouped into three values based on their functionality. For instance, the route category code for U.S. Primary routes, State primary routes, Interstates and Interstate Ramps are grouped as Primary. The Secondary routes, County routes, Urban routes and Ramps are grouped as Secondary. The rest all minor routes are placed into Others category.

Any missing values in Route Category code, Median Type, Toll related variables are filled with ‘UNKNOWN’ value. For structures, the missing values are filled with ‘N’. Filling the missing values helps to handle any future missing value and include it the regression analysis. If there are more than 10% of missing values, the source of the data has to be corrected and these values should be made available. Else, this would cause a greater impact on the analysis.

As the linear regression methodology requires all the explanatory variables to be in numerical format, the categorical variables are converted into dummy variables. The number of dummy variables that were created was one less than number of values under a categorical variable. In route category code, the number of dummy variables would be 3 as it contains four different values (Primary, Secondary, Others and Unknown).

These dummy variables are generated using the python library named pandas. New column names are also assigned to these dummy variables to identify them easily in the analysis.

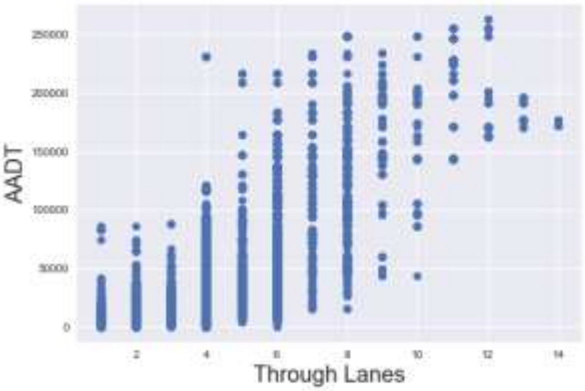

**Fig 1: Variation of AADT with through lanes**

As shown in figure 1, the AADT values increases with the number of lanes. More vehicles can travel through routes with more number of lanes and hence the increases. It can be observed that few points have same level of AADT across different number of lanes. Here, other factors like type of route, population around the route plays a role.

In the 3D plot shown in figure 2, even though the lane counts are higher more than 10, the traffic volume is at the lower range. These are only arterial roads, whereas interstates have higher traffic volume with the same lane count.

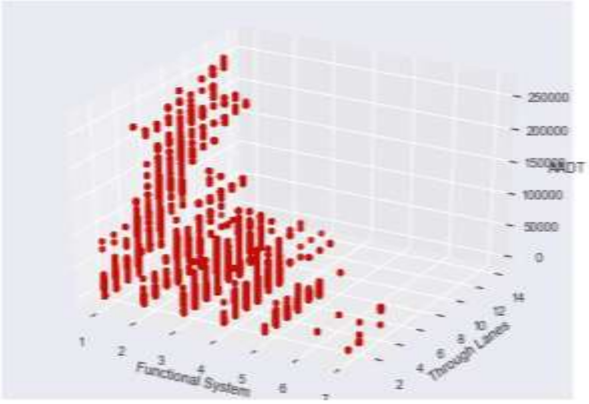

**Fig 2: 3D plot of Functional system, Through lanes and AADT**

## IV. REGRESSION ANALYSIS

### *A. Linear Regression*

Linear regression is one of the useful methods to perform the data analysis and predictions. It helps to interpret the relation between the single or multiple explanatory and a dependent variable. If 'Y' represents the dependent variable and X1…Xn represents the various explanatory variables, then the equation to predict the value of Y would be:

$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots. + bn\ Xn$, where $b_0$ is the slope and b1…bn are the coefficients of the X values.

Here, the Average Annual Daily Traffic is the dependent variable and the explanatory variables are Facility Type, Through Lanes, Functional Class, Route Category code and Toll related fields.

The data is ready to fit into a plot and analyze further. The linear regression equation is plotted used the sklearn library in python. The data is randomly split into train dataset (80%) and test dataset (20%). The train dataset is used to build the linear regression model with the required variables. The model is then evaluated with the test dataset. This also helps to avoid overfitting the model and also to get the accuracy of the predicted model.

The model is created with the train dataset and the values are predicted with the same dataset. The plot in figure 3 shows the variation between the actual and predicted values.

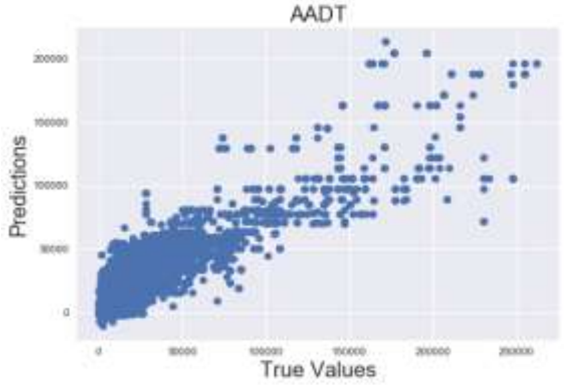

**Fig 3: Scatter plot between True and Predicted values**

The accuracy of the model is calculated with the R2 square value and it came to be 77% accurate.

### *B. Ridge Regression:*

Ridge regression is used when there is multicollinearity in the dataset and it takes to consideration which explanatory variable is important. The ridge regression is similar to the least squares regression but it attempts to move the estimated coefficients towards zero. It is expressed as:

$\beta$^ridge = $(X^T X + \lambda I)^{-1} X^T y$, where I is the Identity matrix and the $\lambda$ parameter is the regularization penalty.

The same dataset as the linear regression is used, but Log 10 is applied for the AADT to see the variation in data clearly. The dataset is split into train and test dataset and the model is built with the Ridge regression method

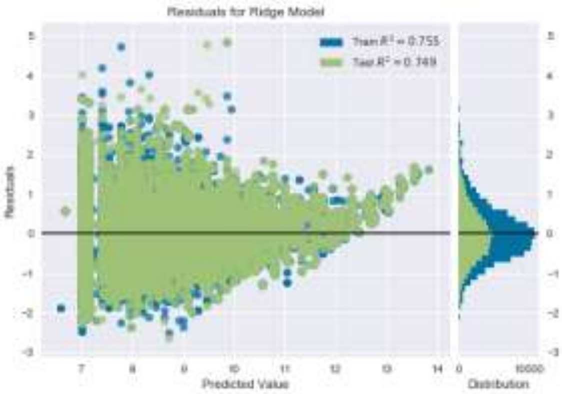

**Fig 4: Residual plot for the Ridge regression model**

The accuracy of the model built with the ridge regression is 76.2 %. In figure 4, the predicted values and the residuals are plotted for both train and test dataset. The distribution of the residuals is shown as a histogram on the right side. The values are normally distributed and have one peak value close to zero.

### *C. Lasso Regression*

The Lasso regression method is similar to the Ridge regression method, which also attempts to shrink the coefficients value towards zero, but it uses the sum of their absolute values to penalize. It is expressed as:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|.$$

In the above expression, the λ is the tuning parameter. It controls the amount of shrinkage on the absolute values. The bias is directly proportional to λ values, whereas the variance is inversely proportional.

The regression is plotted without any transformation on the AADT value, as available in the source. The python libraries sklearn and yellowbrick are been used. In figure 5, the plots between the actual and predicted values are shown. The accuracy percentage turns out to be 76.9%. The plot shows both the identity and best fit lines.

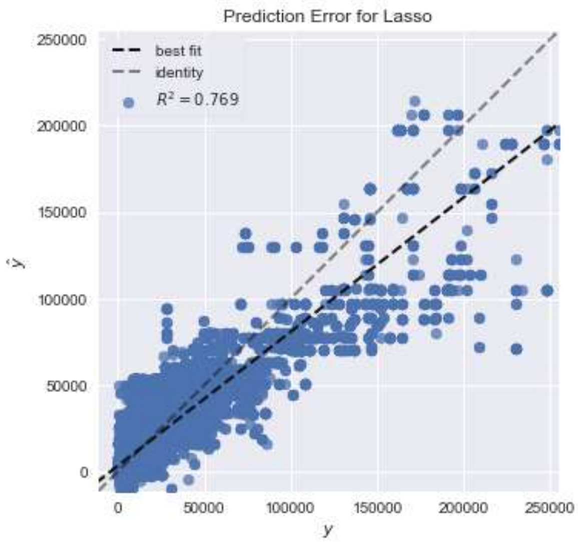

**Fig 5:Prediction plot for Lasso regression method**

These models can be pickled in python, exported and reused for predicting values. When a new route is planned, by providing the explanatory variables, the model can predict the expected traffic volume on the route. Based on the optimal required traffic volume, parameters like through lanes, Toll type can be modified to control the traffic volume. These models have to be revised periodically based on the recent data and this would improve the accuracy of the predictions in future as well.

## V. CONCLUSION

The prediction of AADT values is useful to plan new routes or to extent current routes. With machine learning techniques, these values can be predicted accurately. The Linear regression, Ridge regression and Lasso regression methods were used to predict the AADT value for the State of Virginia. The accuracy between each of the models was almost same. The accuracy can be improved by including other explanatory variables like Surface width, Type of Median, Population count and Access control. Even through urban area code is used in the analysis, actual population count can improve the predicted results.

## REFERENCES

[1] Oleszak, M. (2019, November 12). Regularization: Ridge, Lasso and Elastic Net. Retrieved from https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

[2] Annapoorani Anantharaman,(2019). A Study of Logistic Regression And Its Optimization Techniques Using Octave. SSRG International Journal of Computer Science and Engineering 6(10), 23-28.

[3] Selby, B. &Kockelman, K. M. (2011, January). Spatial prediction of AADT in unmeasured locations by universal kriging.Retrieved from https://www.caee.utexas.edu/prof/kockelman/public_html/TRB11AADTUnivKriging.pdf

[4] Sliupas, T. (2006). Annual average daily traffic forecasting using different techniques. Transport, 21:1, 38-43, doi: 10.1080/16484142.2006.9638039

[5] U. S. Department of Transportation (2018). Traffic Data Computation Method (FHWA-PL-18-027). Retrieved from

[6] https://www.fhwa.dot.gov/policyinformation/pubs/pl18027_traffic_data_pocket_guide.pdf

[7] U. S. Department of Transportation (2018). HPMS Public Release of Geospatial Data in Shapefile Format. Retrievedfrom https://www.fhwa.dot.gov/policyinformation/hpms/shapefiles.cfm

[8] U. S. Department of Transportation (2019). Cost Effective Strategies for Estimating Statewide AADT (FHWA-SC-18-10). Retrieved from https://www.scdot.scltap.org/wp-content/uploads/2019/04/SPR-717-Final-Report.pdf

[9] Virginia Department of Transportation (2019). LRS Route Overlap. Retrieved from https://www.virginiaroads.org/datasets/lrs-route-overlap

[10] Wu, J. & Xu, H. (2019). Annual Average Daily Traffic Prediction Model for Minor Roads at Intersections. Journal of Transportation Engineering, 145(10), doi:10.1061/JTEPBS.0000262

[11] Peixerio, M. (2019, January 12). How to Perform Lasso and Ridge Regression in Python. Retrieved from https://towardsdatascience.com/how-to-perform-lasso-and-ridge-regression-in-python-3b3b75541ad8

[12] Regression Visualizers. (n. d.). Retrieved from https://www.scikit-yb.org/en/latest/api/regressor/

[13] Shakti Chourasiya, Suvrat Jain,(2019). A Study Review On Supervised Machine Learning Algorithms. SSRG International Journal of Computer Science and Engineering 6(8), 16-2