DRISI

CALTRANS DIVISION OF RESEARCH, INNOVATION AND SYSTEM INFORMATION

TRANSFORMING IDEAS INTO SOLUTIONS

Research Notes

A Machine Learning Model for Generating Traffic Data on California Highway System

This project will develop a Machine Learning (ML) model to predict statewide traffic volumes and speeds.

Traffic Operations

MAY 2026

Project Title: A Machine Learning Model for Generating Traffic Data on California Highway System

Task Number: 4413

Start Date: June 1, 2024

Completion Date: November 30, 2026

Task Manager:

Charina Guarino
Transportation Engineer (Civil)
Charina.Guarino@dot.ca.gov

DRISI Division Chief:

Prakash Sah, PE
Prakash.Sah@dot.ca.gov

WHAT IS THE NEED?

The California Department of Transportation (Caltrans) is looking to increase the accuracy and reliability of the traffic data collected statewide. The primary methods of traffic data collection by Caltrans are the use of Intelligent Transportation System (ITS) elements such as Traffic Census and Performance Measurement System (PeMS) stations. The data collected through these methods are limited by the sparsity of covered area, unreliable sensor functionality and unreliable quality. Caltrans needs accurate traffic data to perform essential functions such as traffic flow optimization, infrastructure development, safety enhancement and emergency response. There is a need to improve and ensure Caltrans traffic data is robust and reliable using existing traffic data sources. A ML model would be developed to fill in the gap in Caltrans traffic data. This approach has the potential to eliminate the need to procure additional traffic sensors and other data sources.

WHAT ARE WE DOING?

In this research, the Advanced Highway Maintenance and Construction Technology (AHMCT) Research Center at the University of California (UC) Davis will conduct a literature review on a mixture of data integration techniques including Ensemble Learning, Deep Learning, Time Series Analysis for Temporal Data, Geospatial Analysis, Feature Engineering, Transfer Learning, Outlier Detection, and Multi-Modal Models to determine which technique(s) would be the most appropriate for Caltrans needs. AHMCT will also conduct data preprocessing, exploratory data analysis (EDA) and data integration using traffic data provided by Caltrans. AHMCT will propose features to include in the development

Caltrans

DRISI provides solutions and knowledge that improves California's transportation system.

of ML model for forecasting traffic flow and speed. This will be accomplished in partnership with the Division of Traffic Operation's Business Intelligence & Automation team to understand their needs. The results will be compiled in a final report, along with the ML model and technical documentation for model training and evaluation.

WHAT IS OUR GOAL?

The goal of this research project is to develop a ML model with the ability to forecast and impute traffic volume and speed information, filling in the data gap of current Caltrans traffic data sources.

WHAT IS THE BENEFIT?

The results of this project will help improve the reliability of traffic data available to Caltrans without the procurement of additional data collection equipment and software. The ML model will bridge the data gap across the State Highway System as Caltrans works to meet its Strategic Management Plan.

WHAT IS THE PROGRESS TO DATE?

A 6-month no-cost time amendment was approved extending the project timeline to November 30, 2026.

Project panel meetings were held on August 7, October 17, and December 8, 2025. Contractor designed a hybrid integration pipeline using PeMS sensor data and aligned sensors spatially and temporally to ensure each location's reading were synchronized and normalized.

Contractor developed and applied a sensor anomaly detection framework which identified 5 major sensor anomaly types: flatline, spike, extreme low, stuck high and illogical feature combinations. Contractor conducted quality control of data and established an automated data-validation and fault-detection framework to ensure high-confidence inputs for modeling.

Project panel meetings were held on January 30 and April 29, 2026. Contractor finalized a structured feature pool from the integrated District 4 dataset. Features include traffic and temporal variables, historical lag features, and geographic and roadway metadata. Contractor completed comprehensive benchmarking study across multiple model types including statistical baseline models, tree-based models, and deep learning models. Contractor fully trained three (3) models, XGBoost, LightGBM and STGCN, on a full year of data.