USING ALTAIR® RAPIDMINER® TO ESTIMATE AND VISUALIZE ELECTRIC VEHICLE ADOPTION

INTRODUCTION

Data drives vital elements of our society, and the ability to capture, interpret, and leverage critical data is one of Altair's core differentiators. While Altair's data analytics tools are applied to complex problems involving manufacturing efficiency, product design, process automation, and securities trading, they're also useful in a variety of more common business intelligence applications, too.

An Altair team undertook a project utilizing Altair machine learning (ML) software and data visualization tools to investigate a newsworthy topic of interest today: the adoption level of electric vehicles, including both BEVs and PHEVs, in the United States at the county level.

This guide explains the team's findings and the process they used to arrive at their conclusions.

“Machine learning algorithms can estimate the unknown values in an incomplete dataset. The process of building and using such AI models lets analysts identify variables with high predictive power, and use those indicators to inform future decisions.”

Use Machine Learning Models to Impute Missing Values in Incomplete Datasets

Datasets are rarely perfect and complete in the real world, but thankfully, it's possible to build ML models that can estimate unknown values. This guide explains a particular use case involving analysis of socioeconomic and related data, but the same principles apply in many other applications.

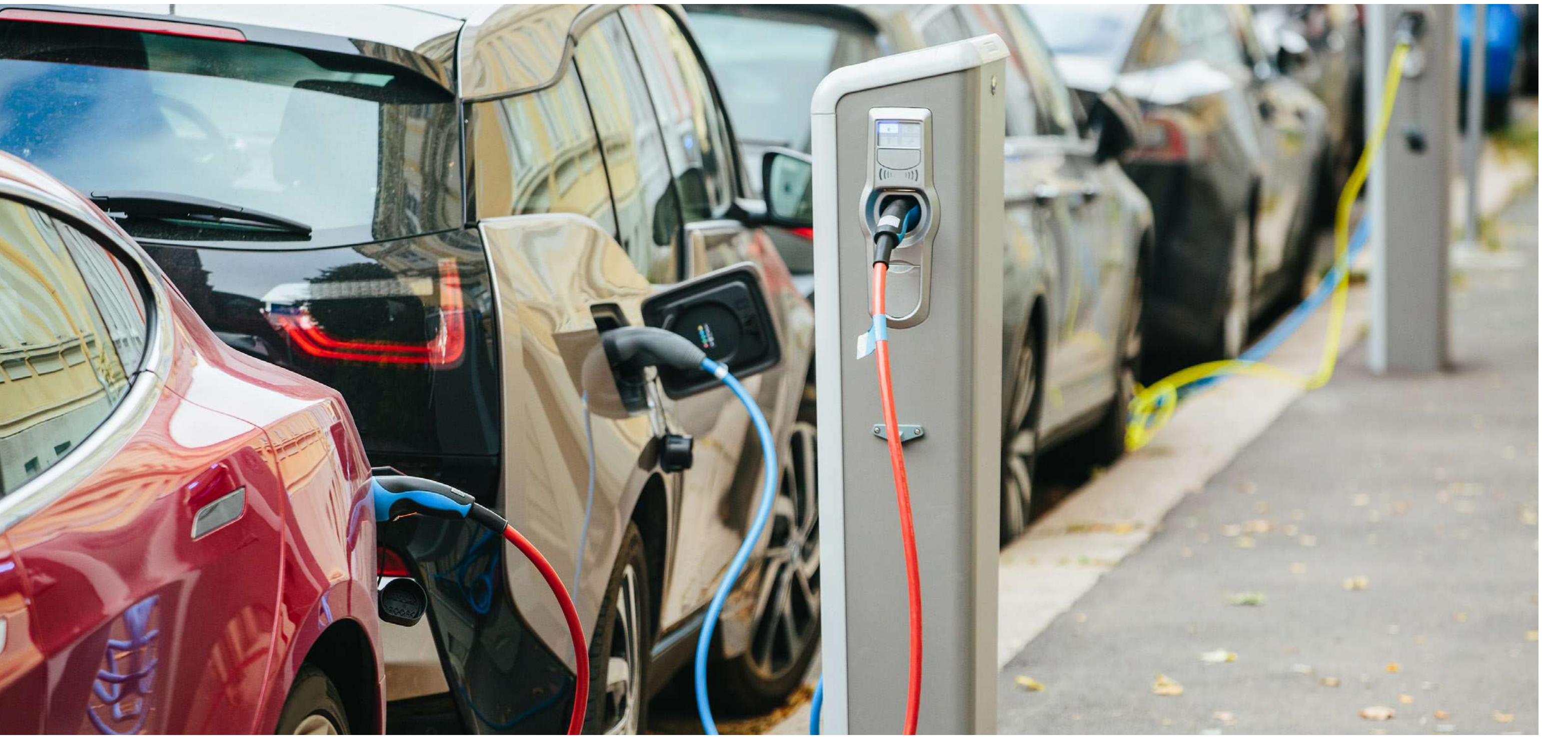
What to Expect

Table of Figures

ESTIMATING UNKNOWN VALUES

Generally, predicting the future garners the most attention in the field of predictive analytics. However, it's also possible to predict the past in the sense of fleshing out incomplete information. For example, the U.S. Food & Drug Administration (FDA) has used large amounts of electronic health records to work out effective various drug treatment regimens. Altair explored this concept in the realm of materials testing. There is an enormous variety of polymer materials that product designers can choose from, with more always entering the market. Altair's data science team has shown how applying the right ML algorithms to documented test results for thousands of existing materials can accurately predict how new, untested materials will perform. Essentially, the concept of predicting the past is an approach for imputing missing values within an existing dataset.

Predicting past unknowns showcases the power of ML and Altair software in making accurate predictions sourced from large amounts of data. By building and applying ML algorithms to incomplete historical data, the team can create complete datasets. This process lets analysts identify variables with high predictive power and then use those indicators to inform future decisions.



Data Sources

The Altair team gathered data from a variety of open sources on vehicle registrations for 15 states at the county level, along with comprehensive socioeconomic, election, population density, and infrastructure data for all 50 states. The project aimed to predict EV adoption level for all counties in the 35 states for which EV registration data wasn't publicly available.

Data sources included:

- Atlas Public Policy EV Hub
- U.S. Department of Energy's Alternative Fuels Data Center
- U.S. Census Bureau
- U.S. Energy Information Administration
- U.S. Department of Agriculture
- Wikipedia
- Homeland Infrastructure Foundation Level Database (HIFLD)

The team found reliable county-level EV registration data for the following 15 states:

- California
- Oregon
- Washington
- Montana
- Colorado
- Texas
- Florida
- Minnesota
- Wisconsin
- Michigan
- New York
- New Jersey
- Vermont
- Tennessee
- Virginia

The team found reliable county-level data for all counties in all 50 states for the following variables:

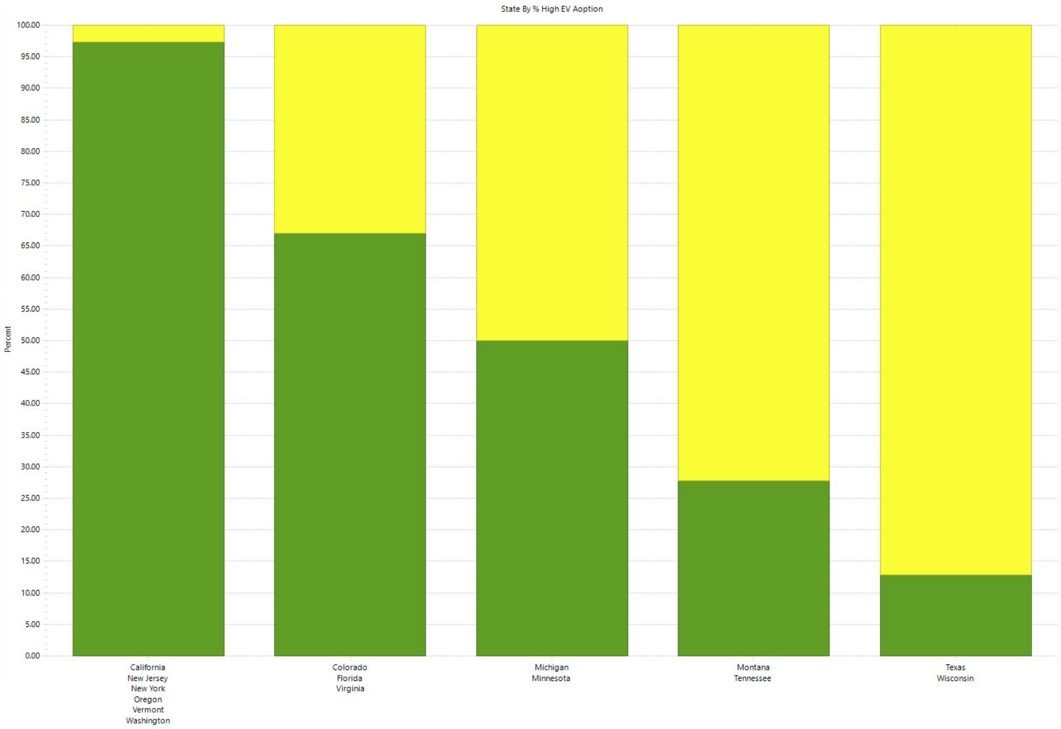
- 2016 presidential election results
- Age
- Charging station availability
- Climate zone classification
- Educational attainment
- Electrical substations
- Family size
- Gender
- Household size
- Households with computers
- Households with internet
- Laws and incentives
- Mean travel time to work
- Means of commute to work
- Median earnings for workers
- Median family income
- Median gross rent
- Median household income
- Median non-family income
- Median rooms in housing units
- Number of vehicles
- Rural-urban classification

The final dataset for the 15 states for which EV registration data was available contains 1,141 records. The dataset for the remaining 35 states contains 2,001 records.

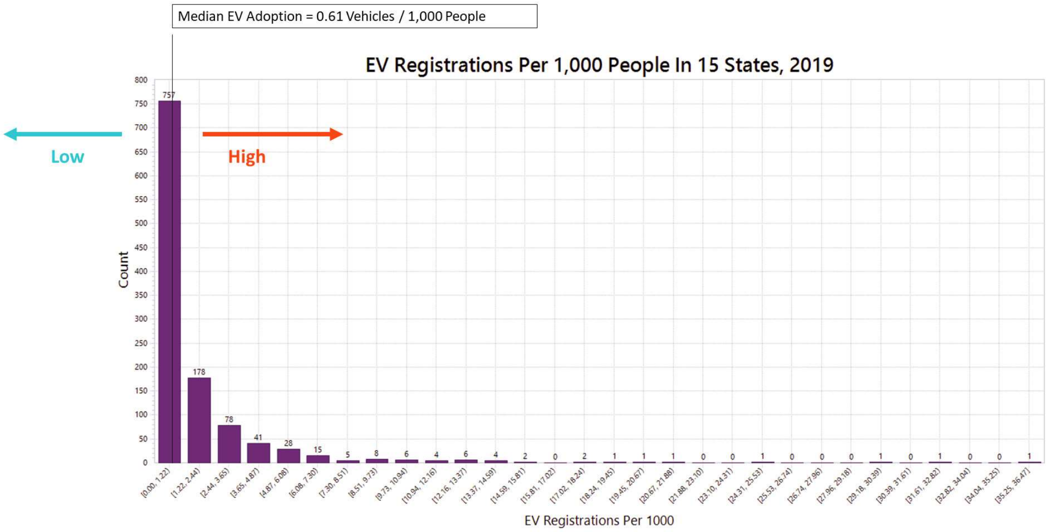
Data Exploration and Visualization

The team brought the predictive modeling data into the [Panopticon data visualization platform] to better understand the known values and relationships within that set. They found that in 2019, California, Washington, Oregon, New York, New Jersey, and Vermont had the highest percentage of counties with high EV adoption rates, while Texas and Wisconsin were states with the lowest percentage of counties with high adoption rates.

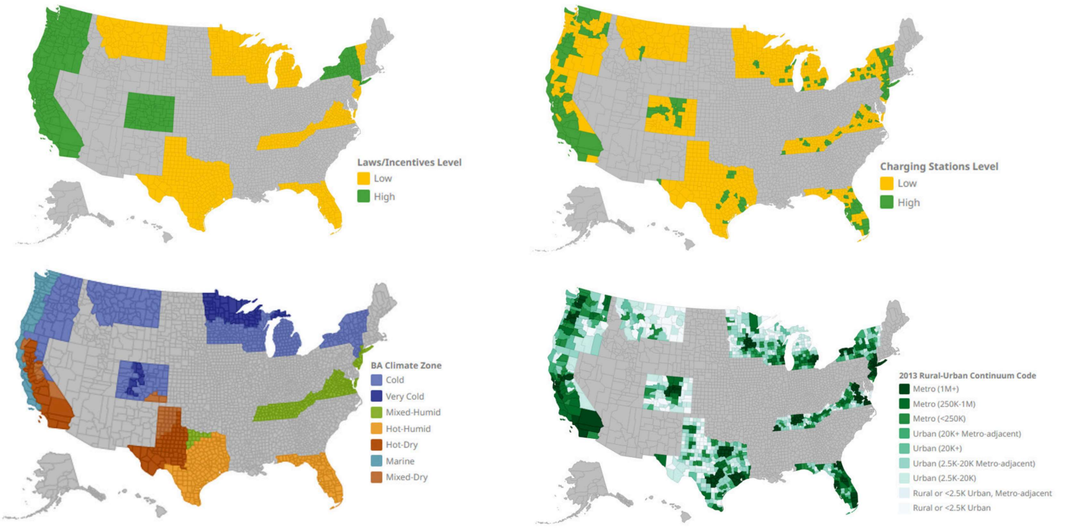
This chart groups states together that share similar percentages of high adoption level counties. The team used an optimal binning feature to create the groupings.



EV adoption level within known states.



The team found that the median EV adoption level in the states with known registration data is 0.61 EVs per 1,000 people.



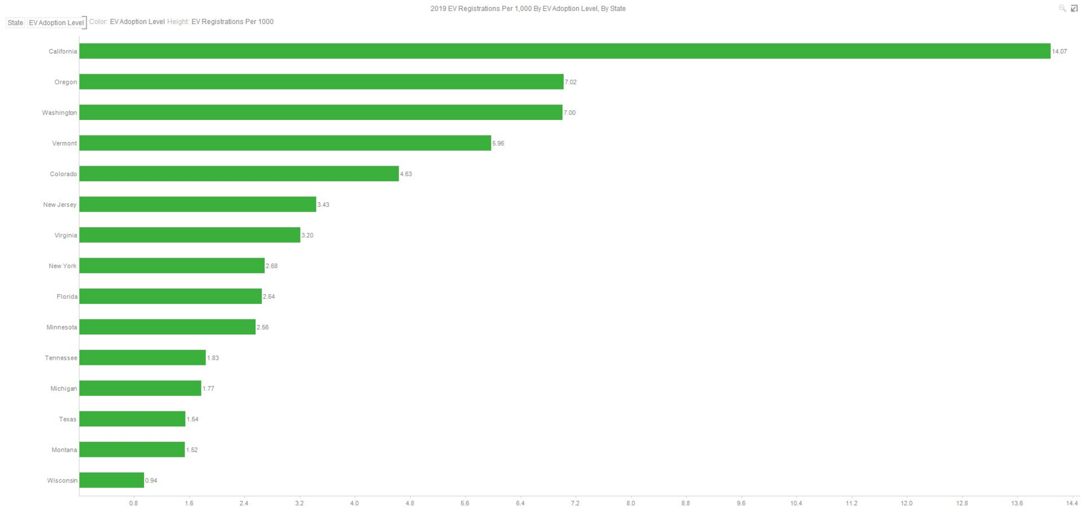
The team used Panopticon to create choropleth maps visualizing laws and incentives, charging station, climate zones, and census data by county for the 15 states with known vehicle registration data.

adoption rates:

- Average number of vehicles per household
- Average commute time
- Age distribution

The team determined that counties meeting the following criteria were most likely to have high adoption level:

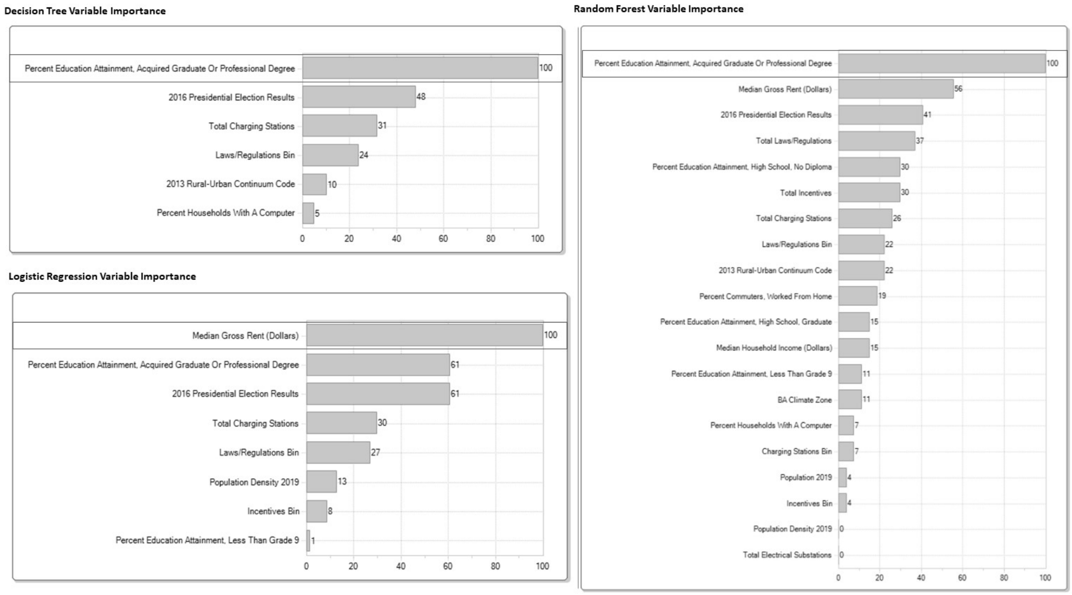
- Median gross rent of $900 or more
- Median household income of $65,000 or more
- 10% or more of the county population over the age of 25 has a graduate degree
- 20% or more of the county population over the age of 25 has a bachelor's degree
- 80% or more of households have active broadband internet subscriptions
- 90% or more of households have at least one computer
- The county is in a metropolitan or populous urban area
- The county is in a state that voted for Hillary Clinton in 2016



The bar chart was created using the Panopticon data visualization platform and displays the average number of EVs per 1,000 population registered in the 15 states in the training dataset.

PREDICTIVE MODELING

The team looked at county-level demographic and socioeconomic data for the 15 states for which they had EV registration data. Figure 5 displays the results of how three different algorithms determined which variables contain the highest predictive power of EV adoption rates.



Decision Tree Variable Importance

Logistic Regression Variable Importance

Random Forest Variable Importance

The analytics team used Knowledge Studio to develop ML models, including decision tree, logistic regression, and random forest algorithms to predict EV adoption level in counties in which data was unavailable.

The team found that the following variables held high predictive power for EV adoption level at the county level:

- Percent of population with graduate or professional degrees
- Median gross rent
- 2016 presidential election results
- Number of charging stations
- Population density

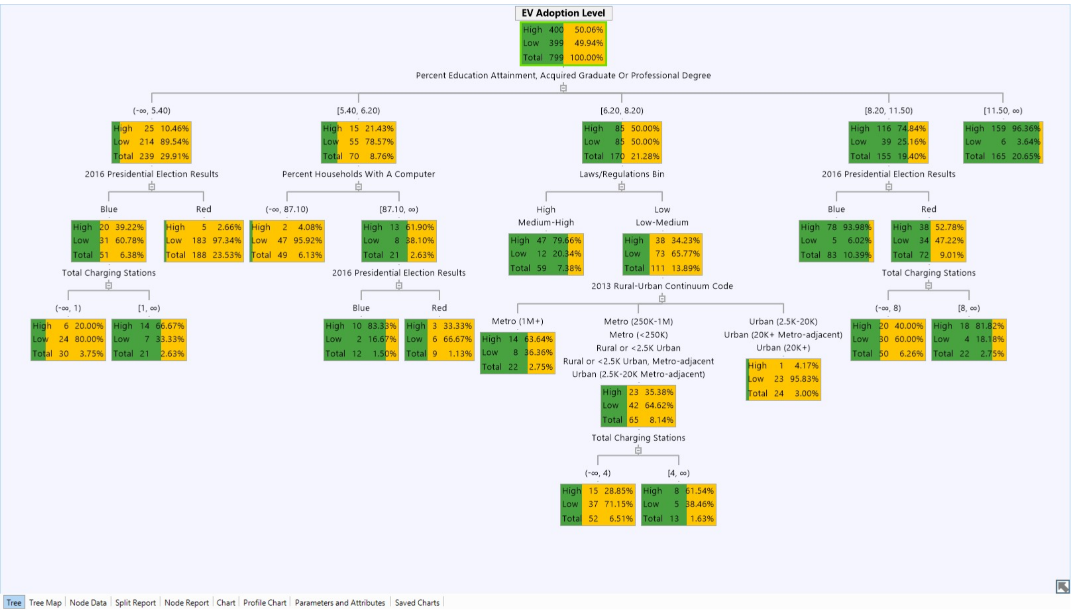
Percent of population with graduate or professional degrees

Median gross rent

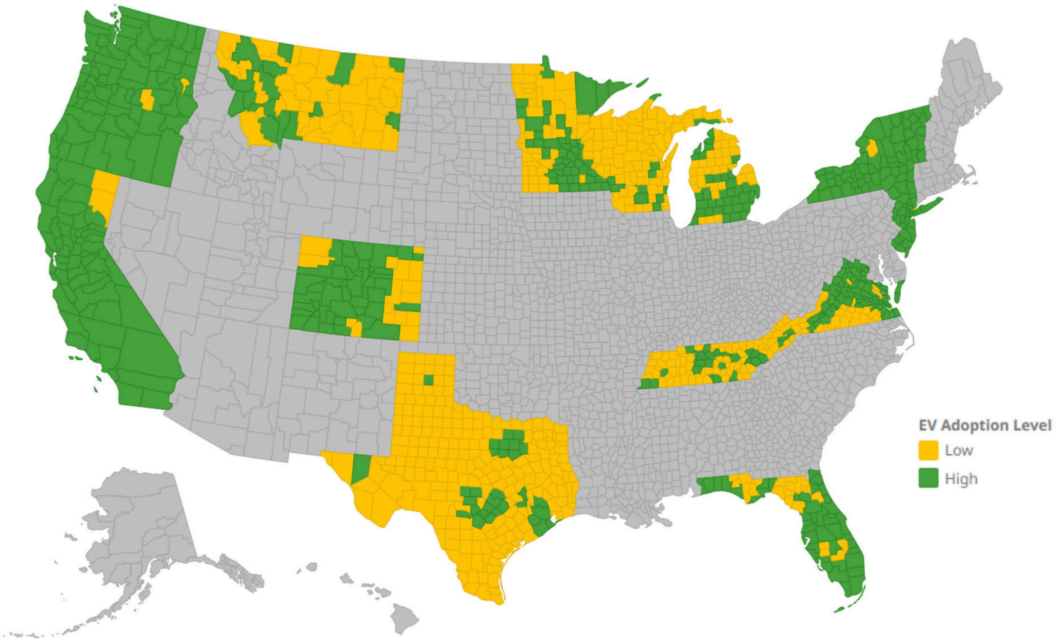
2016 presidential election results

Number of charging stations

Population density



Knowledge Studio decision trees identify relationships in subgroups of data and generate predictions from the characteristics of other datasets for which EV registration data is unavailable.

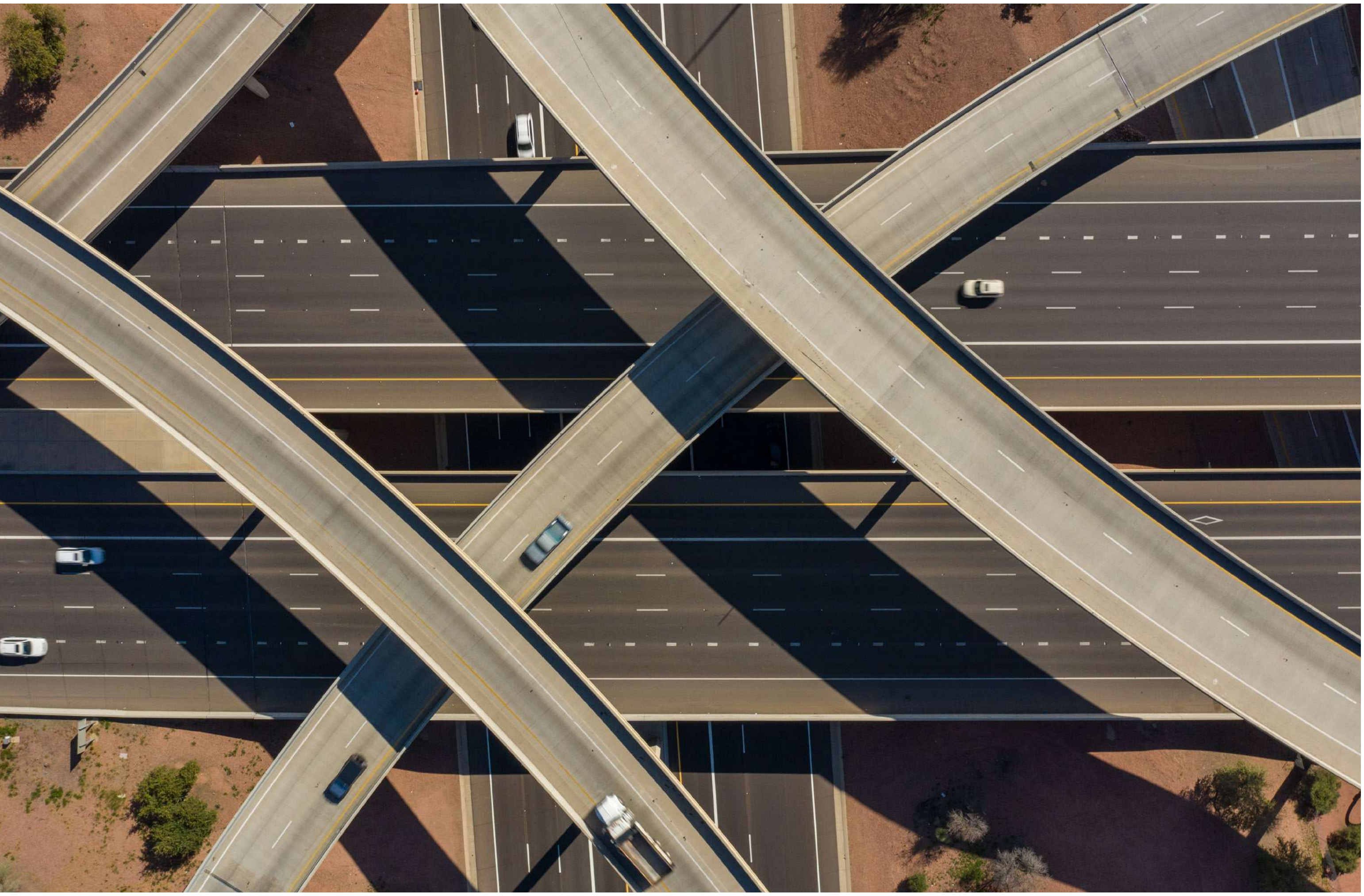


This choropleth plot shows the EV adoption level at the county level for the 15 states with known registration data.

MODEL VALIDATION

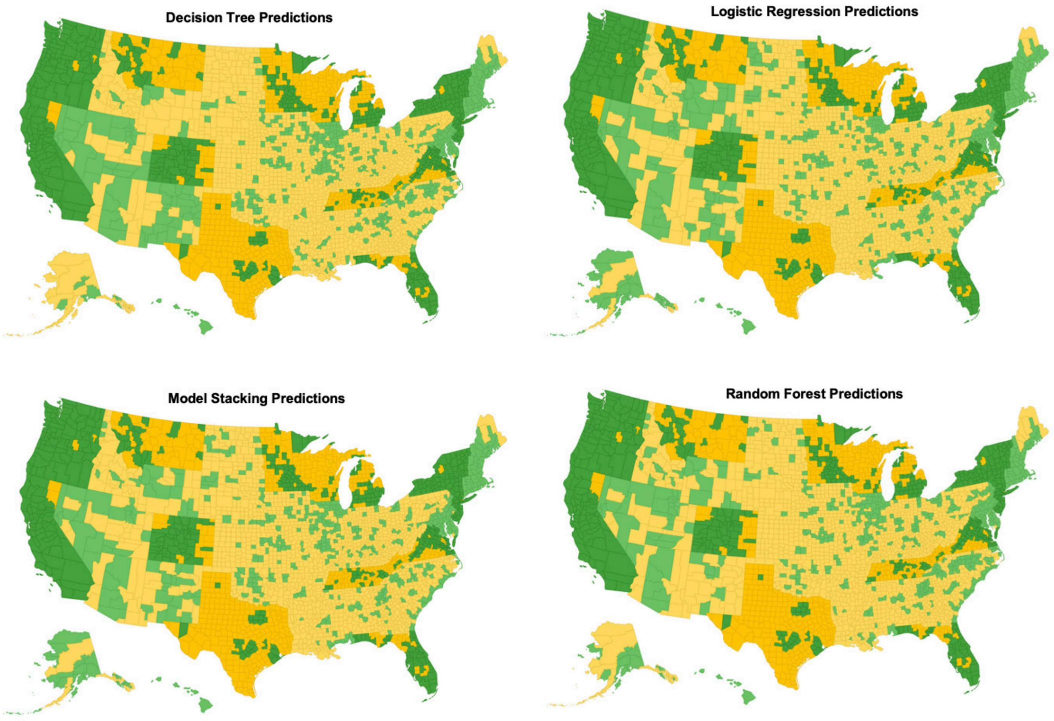
The team used three primary methods to validate their predictions:

- Comparing their predictions with EV market share data from the International Council on Clean Transportation
- Dividing the predictive modeling dataset for the 15 known states into two subsets: training and test. They used ML models to make predictions for the test data subset and compared the predictions with actual data recorded in the training data subset.
- Building three different predictive models based on random forest, logistic regression, and decision tree algorithms. They then compared the predictions made by each model and found only small variances. The team also used a model stacking technique that combines all three models in an attempt to achieve a higher level of accuracy to see if they could get better performance. After examining the results of these comparisons, the team found that the random forest-based model was the most accurate.

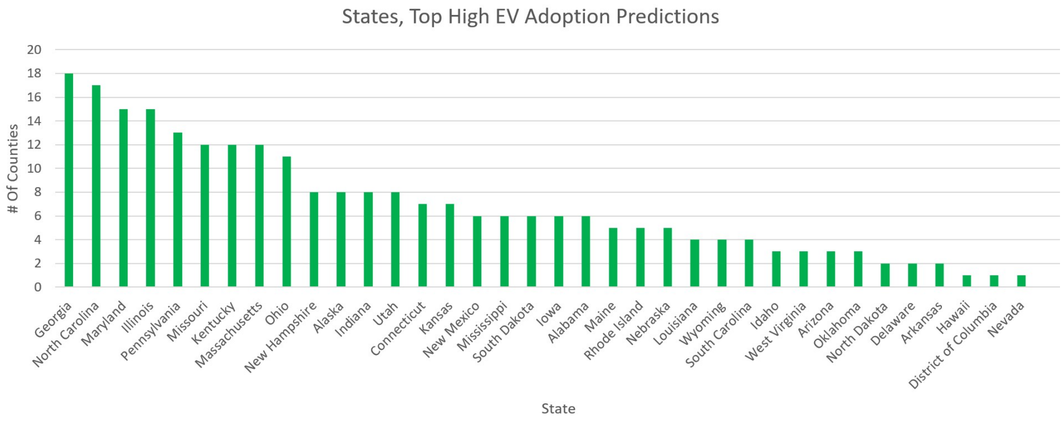


RESULTS

The predictions led to some surprising results. New Mexico, Arizona, and Nevada, despite being large, mostly rural states with long travel distances, will adopt EVs at a level like more urbanized states like Florida. Also, there are numerous adoption hotspots in almost every state, including those in the Deep South and the Midwest.


Decision Tree Predictions
Logistic Regression Predictions
Model Stacking Predictions
Random Forest Predictions

These choropleth plots show predicted EV adoption levels for all counties in the U.S., including the team's predictions for the 35 states without EV registration data, based on the outputs of four different ML algorithms.



This bar chart shows the states with the most counties likely to have high EV adoption level based on predictions made using a decision tree.

ALTAIR ADDRESSES DATA ANALYTICS CHALLENGES

Insight comes when you can access the hundreds or thousands of dimensions hidden in complex data, including in situations where source data is incomplete. People need the right tools to access those hidden dimensions, build reliable source datasets, and visualize the results. Altair empowers business users to collaborate efficiently and access meaningful data, generate insight from this data, and share their finding throughout the enterprise.

Data Preparation

Altair's data prep software lets businesspeople build, discover, share, and collaborate on secure, governed, and trustworthy data sets and models. These tools can access, cleanse, and format data from a wide variety of sources (including Excel, CSV, PDF, TXT, JSON, XML, HTML, SQL databases, big data like Hadoop, and more) without any manual data entry or coding. Dozens of pre-built data preparation functions make combining disparate but related data sets quick and easy. This simple approach to data cleansing eliminates the need to code, script, or create pivot tables or vLookups in Excel. Clients can deploy these tools on desktop, with on-premise servers, or in the cloud.

Predictive Modeling and Machine Learning

Altair's open, flexible predictive analytics platform is designed for data scientists and business analysts alike. Its industry-leading visual approach to analytic modeling help data science teams create high quality ML and AI models. Our collaborative approach to ML helps your data scientists and business users minimize repetitive tasks, share knowledge across the enterprise, and reuse steps within connected model workflows for faster analysis. Altair's code-optional development environment enables data science teams to build models using combinations of SAS language, Python, R, and SQL code.

Data Preparation

Predictive Modeling and Machine Learning

Stream Processing

Data Visualization

Stream Processing

Altair's stream processing (also referred to as "event processing") engine connects to a range of real-time streaming and historic time series data sources, including MQTT, Kafka, Solace, and many others. Users can build complex stream processing applications with a drag-and-drop interface, without needing to write any code. Applications may combine streaming data with historic data, calculate performance metrics using a variety of statistical and mathematical functions, aggregate, conflate, and compare data sets, and automatically highlight anomalies against user-defined thresholds.

Data Visualization

Altair's visual analytics platform is optimized for handling time-critical data, including data that may be changing rapidly. Business users can connect to data sources, build, and publish sophisticated real-time dashboards. The platform's filtering tools let users zoom in and out on the timeline, remove false positives from the screen, and focus on exceptions. Users can solve difficult problems quickly, understand complex relationships in seconds, and identify issues requiring further investigation with just a few clicks.

Altair is a global leader in computational science and artificial intelligence (AI) that provides software and cloud solutions in simulation, high-performance computing (HPC), data analytics, and AI. Altair enables organizations across all industries to compete more effectively and drive smarter decisions in an increasingly connected world – all while creating a greener, more sustainable future.

For more information, visit www.altair.com