

Machine learning with tidymodels

Regression analysis of electric vehicle ranges

Made by:Hampus Nordholm

2024-10-06

Introduction

The report aims to analyze the average effect on the range in miles based on electric vehicle type (battery electric vehicle or plug-in hybrid). A linear regression model was trained using tidymodels and evaluated on a testing split. Correlation analysis was also performed to identify features with a linear relationship to longer range. Through regression analysis, the main goal is to answer the question:

What is the effect on range in miles for electric vehicles depending on whether they are battery electric vehicles or plug-in hybrids?

Solution summary

The regression model achieved an RMSE of 53.8, an R-squared value of 0.705, and an MAE of 36.1 when evaluated on the testing split. There is a statistically significant difference in range in miles between plug in hybrids and battery electric vehicle types, with an average decrease of 167 miles for plug in hybrids compared to battery electric vehicles. Through correlation analysis, it was concluded that Tesla models such as the Bolt EV and Model 3 are the most strongly correlated with the range bin of 215 to infinity. Chevrolet and Tesla were the only manufacturers positively correlated with the highest range bin within this analysis.

Core syntax for analysis

```
#CORE LIBRARIES
```

```
#Data analysis
```

```
library(tidyverse)
```

```
library(correlationfunnel)
```

```
library(skimr)
```

```
library(janitor)
```

```
#Machine learning
```

```
library(tidymodels)
```

```
#Loading data --
```

```
electric_tbl <- read_csv("data.csv")
```

```
## Rows: 205439 Columns: 17
## -- Column specification -----
## Delimiter: ","
## chr (10): VIN (1-10), County, City, State, Make, Model, E.V_Type, CAFV, Vehi...
## dbl (7): Postal Code, Model Year, Electric Range, Base MSRP, Legislative Di...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Data exploration --
```

```
electric_tbl %>% glimpse()
```

```
## Rows: 205,439
## Columns: 17
## $ `VIN (1-10)`      <chr> "JTMAB3FV3P", "1N4AZ1CP6J", "5YJ3E1EA4L", "1N4A~
## $ County           <chr> "Kitsap", "Kitsap", "King", "King", "Thurston",~
## $ City             <chr> "Seabeck", "Bremerton", "Seattle", "Seattle", "~
## $ State            <chr> "WA", "WA", "WA", "WA", "WA", "WA", "WA", "WA",~
## $ `Postal Code`    <dbl> 98380, 98312, 98101, 98125, 98597, 98036, 98370~
## $ `Model Year`     <dbl> 2023, 2018, 2020, 2014, 2017, 2020, 2022, 2023,~
## $ Make             <chr> "TOYOTA", "NISSAN", "TESLA", "NISSAN", "CHEVROL~
## $ Model            <chr> "RAV4 PRIME", "LEAF", "MODEL 3", "LEAF", "BOLT ~
## $ E.V_Type         <chr> "PHEV", "BEV", "BEV", "BEV", "BEV", "BEV", "PHE~
## $ CAFV             <chr> "known", "known", "known", "known", "known", "k~
## $ `Electric Range` <dbl> 42, 151, 266, 84, 238, 291, 31, 0, 291, 84, 238~
## $ `Base MSRP`      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 31950, 0, 0~
## $ `Legislative District` <dbl> 35, 35, 43, 46, 20, 21, 23, 39, 47, 45, 26, 35,~
## $ `DOL Vehicle ID` <dbl> 240684006, 474183811, 113120017, 108188713, 176~
## $ `Vehicle Location` <chr> "POINT (-122.8728334 47.5798304)", "POINT (-122~
## $ `Electric Utility` <chr> "PUGET SOUND ENERGY INC", "PUGET SOUND ENERGY I~
## $ `2020 Census Tract` <dbl> 53035091301, 53035080700, 53033007302, 53033000~
```

```
electric_tbl %>% sample_n(20)
```

```
## # A tibble: 20 x 17
##   `VIN (1-10)` County   City      State `Postal Code` `Model Year` Make  Model
##   <chr>         <chr>   <chr>    <chr>      <dbl>        <dbl> <chr> <chr>
## 1 1C4JJXP67P   King    Tukwila  WA          98188        2023 JEEP  WRAN~
## 2 5YJSA1H2XF   Benton  Kennewick WA          99337        2015 TESLA MODE~
## 3 7SAYGDEE6P   King    Seatac   WA          98198        2023 TESLA MODE~
## 4 1N4BZ1BVXM   King    Shoreline WA          98177        2021 NISS~ LEAF
## 5 5YJ3E1EB4P   Pierce  Tacoma   WA          98406        2023 TESLA MODE~
## 6 5YJ3E1EB4K   King    Seattle  WA          98125        2019 TESLA MODE~
## 7 1C4JJXP60M   Snohomish Edmonds  WA          98026        2021 JEEP  WRAN~
## 8 5YJ3E1EB3N   King    Burien   WA          98168        2022 TESLA MODE~
## 9 5YJSA1E29K   Clark   Brush Pr~ WA          98606        2019 TESLA MODE~
## 10 1N4BZ1DV7N  Snohomish Bothell  WA          98012        2022 NISS~ LEAF
## 11 JTJHKCFZ8S  Jefferson Port Lud~ WA          98365        2025 LEXUS NX
## 12 5YJ3E1EBXL  King    Sammamish WA          98074        2020 TESLA MODE~
## 13 7SAYGAEE1P   King    Maple Va~ WA          98038        2023 TESLA MODE~
## 14 7SAYGDEE6R   King    Seattle  WA          98119        2024 TESLA MODE~
## 15 7SAYGAEE1N   King    Seattle  WA          98134        2022 TESLA MODE~
```

```
## 16 1N4AZ1CP9J King Seattle WA 98115 2018 NISS~ LEAF
## 17 7FCEHEB73P King Tukwila WA 98168 2023 RIVI~ EDV
## 18 5YJ3E1EC6M Pierce Puyallup WA 98374 2021 TESLA MODE~
## 19 YV4ED3GL9P King Seattle WA 98105 2023 VOLVO C40
## 20 7SAYGAEEOOR King Seattle WA 98105 2024 TESLA MODE~
## # i 9 more variables: E.V_Type <chr>, CAFV <chr>, `Electric Range` <dbl>,
## # `Base MSRP` <dbl>, `Legislative District` <dbl>, `DOL Vehicle ID` <dbl>,
## # `Vehicle Location` <chr>, `Electric Utility` <chr>,
## # `2020 Census Tract` <dbl>
```

```
#Cleaning var names --
electric_tbl <- electric_tbl %>% clean_names()
```

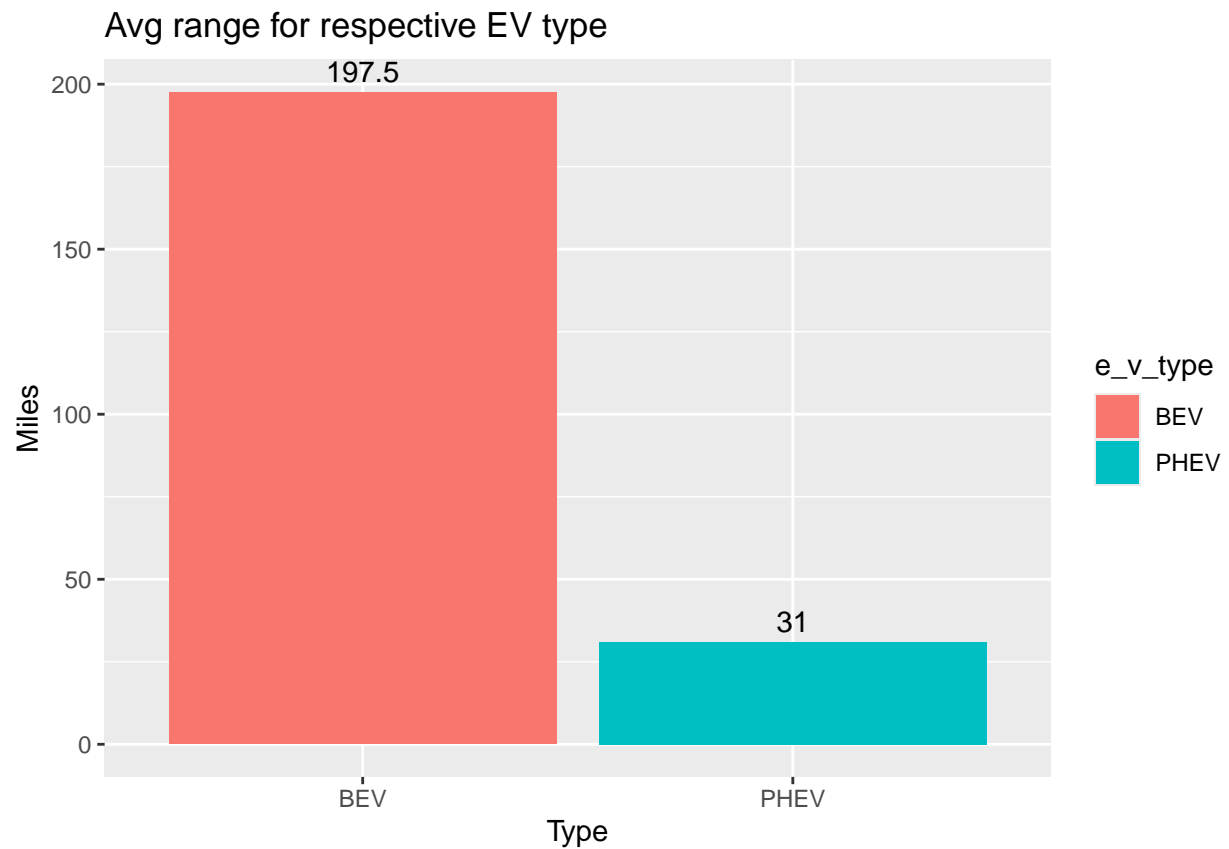
```
#EXPLORATORY DATA ANALYSIS --
```

```
# Count n vehicles where distance = 0 --
electric_tbl %>%
  filter(electric_range==0) %>%
  count()
```

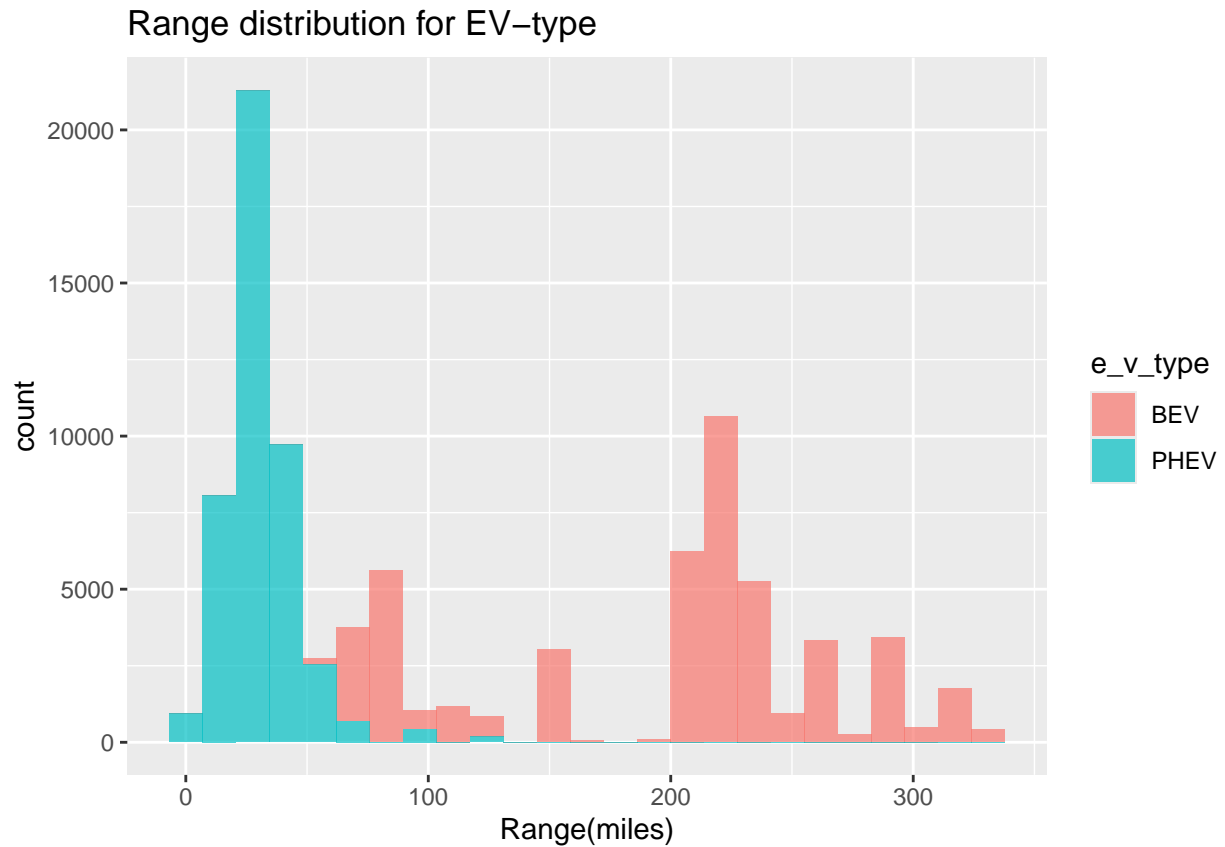
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 114172
```

```
# Basic mean values for electric range by EV type --
```

```
electric_tbl %>%
  filter(electric_range!=0) %>%
  group_by(e_v_type) %>%
  summarise(avg_electric_range=mean(electric_range)) %>%
  ggplot(aes(e_v_type,avg_electric_range,fill=e_v_type))+
  geom_col()+
  geom_text(aes(label=round(avg_electric_range,1)),
            vjust=-0.5)+
  labs(title="Avg range for respective EV type",
       x="Type",y="Miles")
```



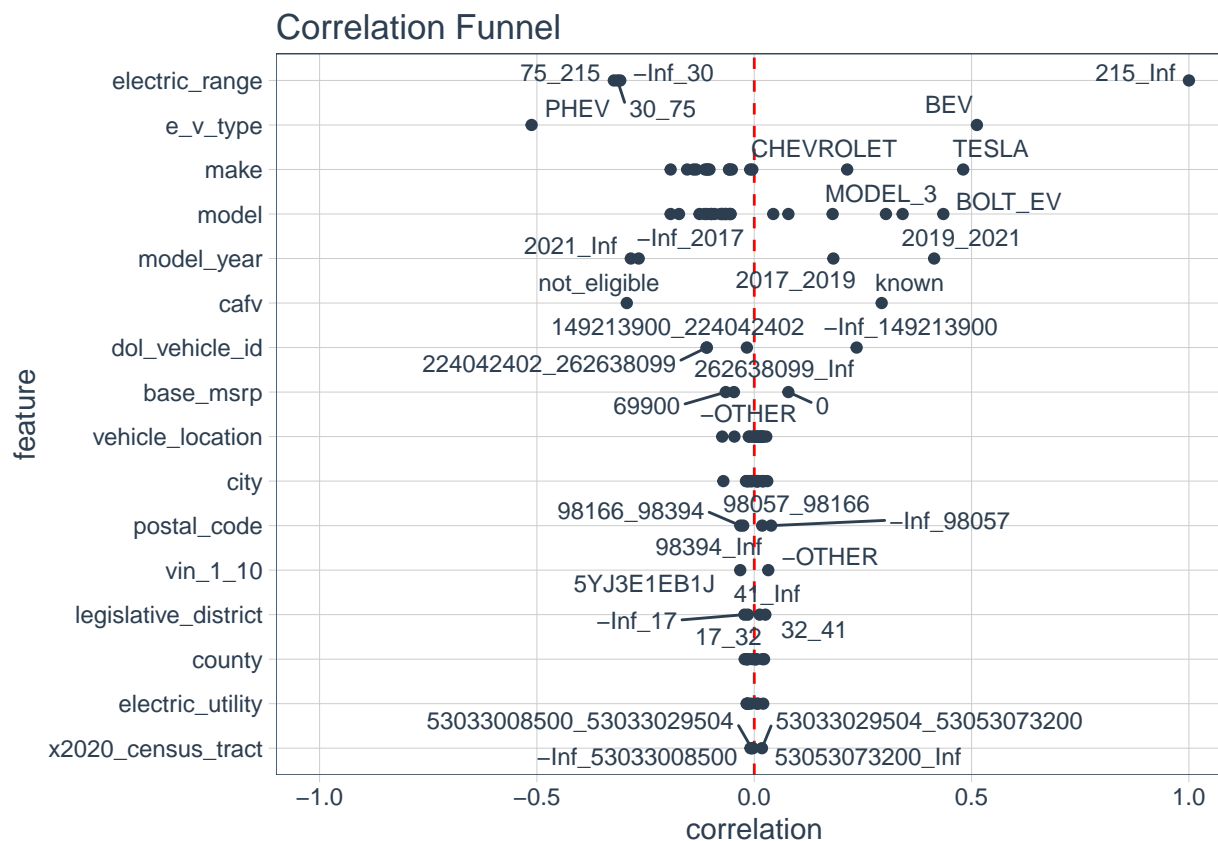
```
#Histogram for respective group --
electric_tbl %>%
  filter(electric_range!=0) %>%
  ggplot(aes(electric_range,fill=e_v_type))+
  geom_histogram(alpha=0.7,bins=25)+
  labs(title="Range distribution for EV-type",
        x="Range(miles)")
```



```
#Correlation analysis --
```

```
electric_tbl %>%  
  filter(electric_range!=0) %>%  
  na.omit() %>%  
  binarize() %>%  
  correlate(electric_range__215_Inf) %>%  
  plot_correlation_funnel()
```

```
## Warning: ggrepel: 96 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps
```



*# Simple linear reg feature selection -- avg. electric range effect of EV-type***

```
ev_type_tbl <- electric_tbl %>%
  filter(electric_range!=0) %>%
  select(electric_range,e_v_type) %>%
  mutate(e_v_type=as.factor(e_v_type))
```

Train / test split

```
set.seed(123)
simple_lm_split <- initial_split(data=ev_type_tbl,prop=0.8)
lm_training <- training(simple_lm_split)
lm_testing <- testing(simple_lm_split)
```

#Regression recipe --

```
lm_model_rec <- recipe(electric_range~e_v_type,data=lm_training)
```

Linear model spec --

```
lm_model_spec<-linear_reg() %>%
  set_engine("lm")
```

#Combine into workflow --

```
lm_wf <- workflow() %>%
  add_recipe(lm_model_rec) %>%
  add_model(lm_model_spec)
```

```
# Training linear model --
```

```
lm_model_fit <- fit(lm_wf,data=lm_training)
```

```
#Results --
```

```
lm_model_fit %>%  
  extract_fit_parsnip() %>%  
  tidy()
```

```
## # A tibble: 2 x 5
```

##	term	estimate	std.error	statistic	p.value
##	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
## 1	(Intercept)	198.	0.275	718.	0
## 2	e_v_typePHEV	-167.	0.397	-420.	0

```
#Model evaluation -- (Examined on testing data)
```

```
ev_predict <- predict(lm_model_fit,new_data=lm_testing)
```

```
#Combining actual vs. predicted values --
```

```
actvspred_lm<- lm_testing %>% select(electric_range) %>%  
  bind_cols(ev_predict)
```

```
lm_evaluation <-metrics(data=actvspred_lm,truth=electric_range,estimate=.pred)
```

```
#Final linear reg. model metrics --
```

```
lm_evaluation
```

```
## # A tibble: 3 x 3
```

##	.metric	.estimator	.estimate
##	<chr>	<chr>	<dbl>
## 1	rmse	standard	53.8
## 2	rsq	standard	0.705
## 3	mae	standard	36.1