

PA Workshop

* SEDANS

**SEDANS**

1. **Data Characteristics and Exploration**

* The variables of total.cost, mileage and vehicle.age are integers and no need to transform after testing the linearity with these continuous independent variables.
* 1489 rows of outliers with standard deviations of the lot.sale.days greater than 2 times of the median are removed.
* There’s no correlation among the numeric variables of total.cost, mileage and vehicle.age and lot.sale.days.

1. **Model Selection, Interpretation & Prediction**

* Upon initial model against all variables, there is no significant linear relationship across all states while there is for individual state. So we assume each state requires a separate model and we select FL state to build a model.
* Upon analysing VIF and relative importance of various predictors, we decided to remove the insignificant variables of vehicle.age.group, make.model and domestic.import. Therefore the remaining variables of our model are total.cost, mileage, vehicle.type, vehicle.age, color.set and makex.
* After trying to include interactions among all combinations of signification variables, we conclude there’s no interaction existing.
* Our final model gets an R square on test dataset close to that of the train dataset.

1. **Business Significance and P&L analysis:**

In general, all variables in the model play significant roles in prediction. While vehicle type, total cost and manufacturer brand show the highest importance. From a business point of view, vehicle type family large, medium and small have high negative correlation with lot sale days, while vehicle age affects the lot sale days in a positive manner.

For specific brands, Cadillac, Chrysler, Honda, Toyota, Dodge, Mazda, Mercury and Pontiac are the top brands to buy as it has significant negative correlation with lot sale days, whereas the sales team should avoid Oldsmobile and Daewoo due to high positive correlation in the model.

After applying the model on validation data set, we found the above business significance are still held. Vehicle age follows a normal distribution and the best purchase point is when the car is between 4 and 7 years old. Also, Family medium type shows popularity as most cars with lot sale days less than 90 comes from the type.

In conclusion, the sales team could focus on cars with vehicle age between 4-7, with family vehicle type. For brand choice, Cadillac, Chrysler, Honda, Toyota, Dodge, Mazda, Mercury and Pontiac could be better than others.