DAT24 Final Project

Dan Lee

The Problem

- A typical car depreciates by $150 a week
- The faster a car sells the larger the profit margin

The Goal

- Predict the amount of time for a car to sell
- Develop an understanding for what features make a car that sell fast (35 days or less)

The Data

- Data consists of internal sales data with customer info removed
- Removed columns with null features and replaced values for columns that had some nulls
- Created additional features such as age group, primary car class, and secondary car class

```
distance_from_dealer
sales_source
free_shipping_cost
stock_number
car_class
body
color
year
make
model
mileage
age
trade_acv
acquisition_cost
reconditioning_cost
total_cost
sale_price
front_gross
financing
warranty
insurance
other_products
back_gross
total_gross
```

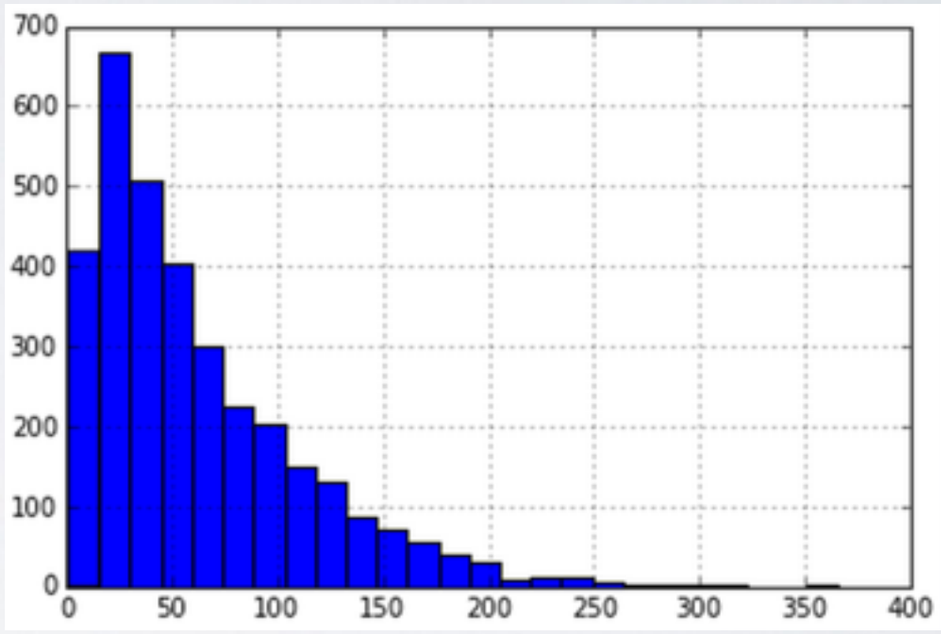
Feature Creation

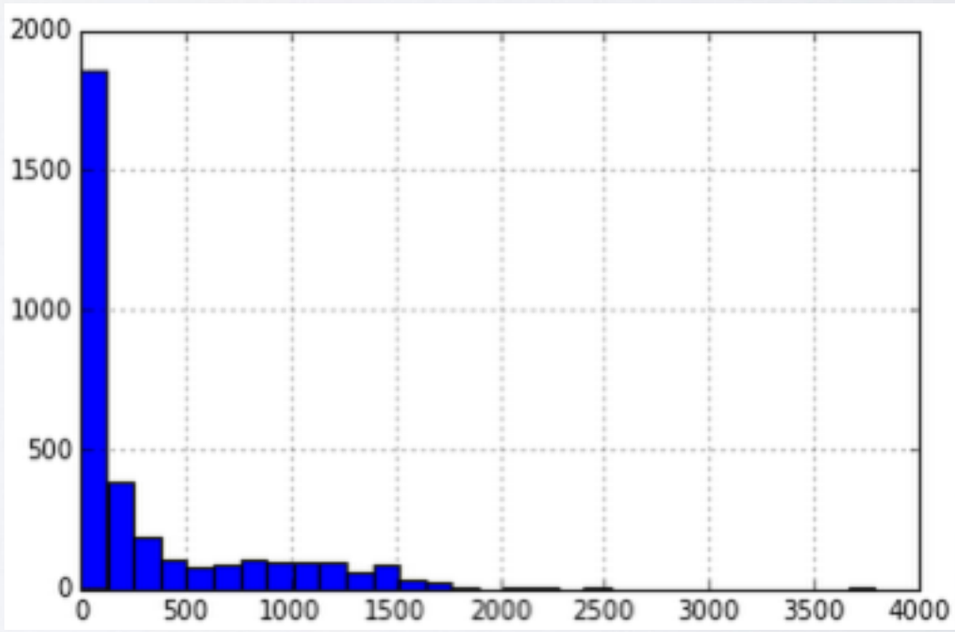
- Determined age groups based on the distribution of the age of historical sales

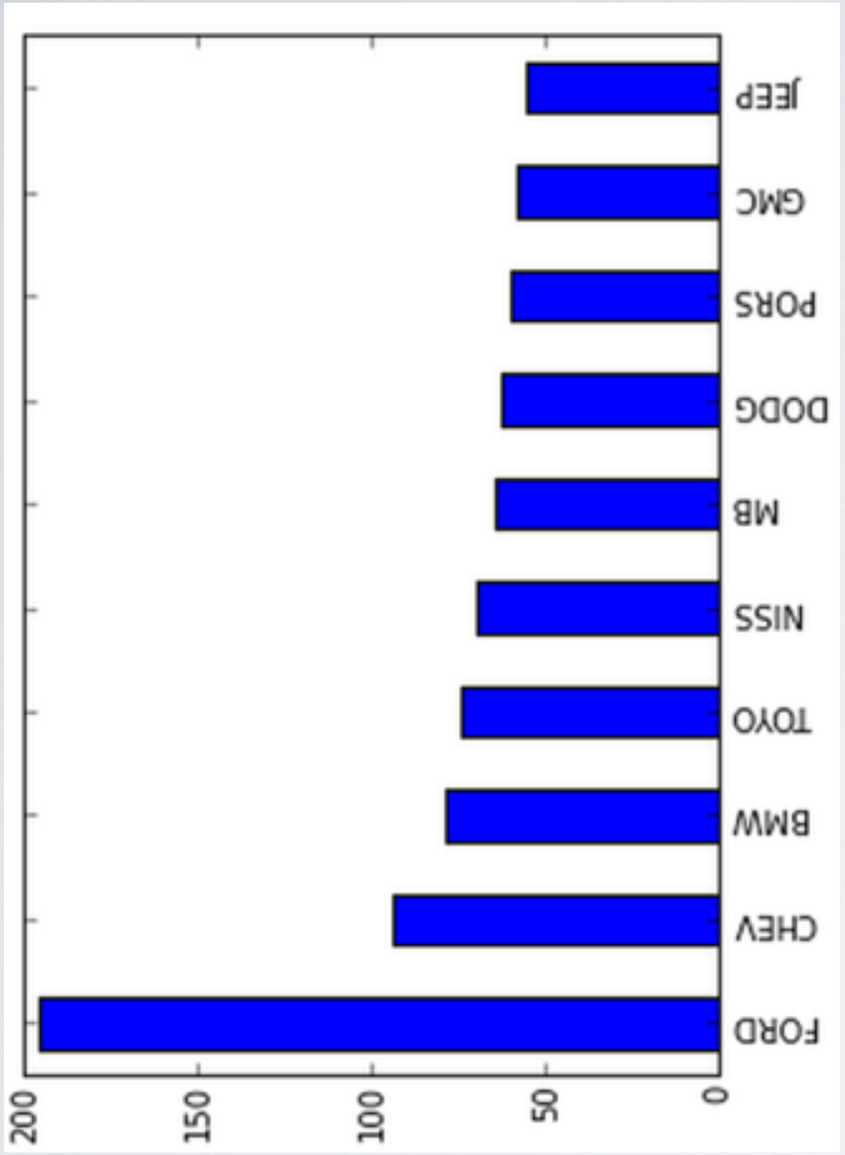
- Distance from dealer was grouped in a similar way

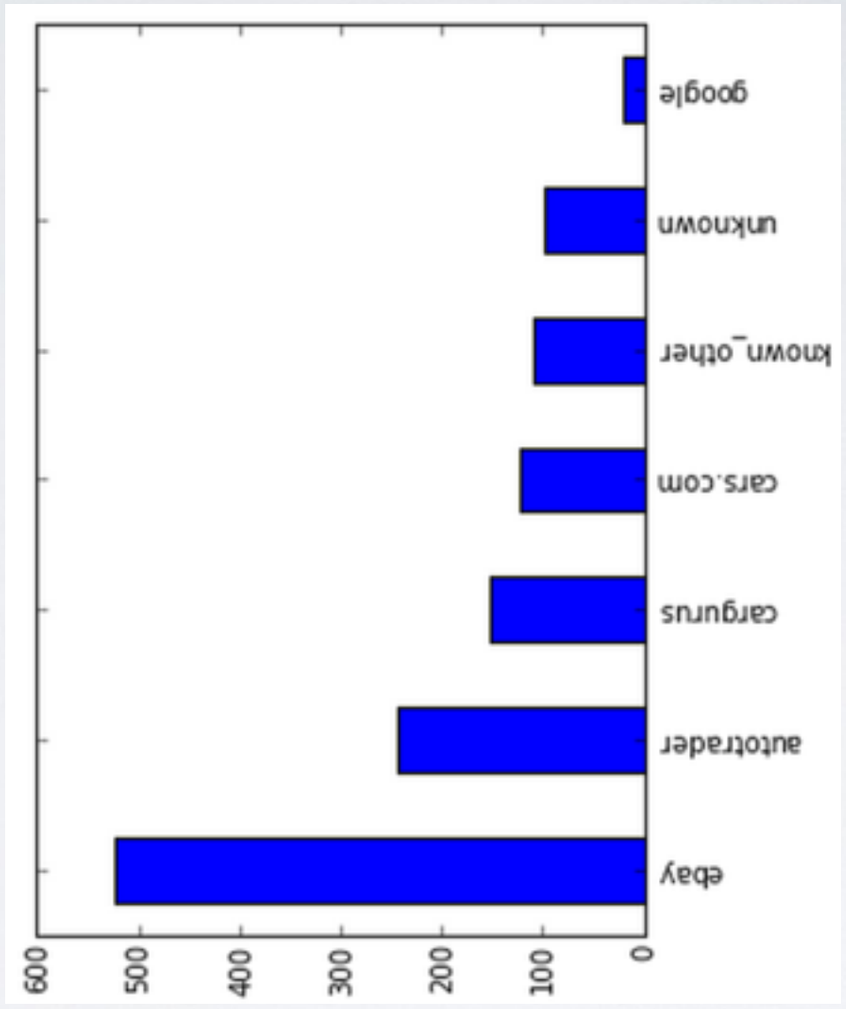
- Broke out car class into 2 features
  - Primary and Secondary car class.

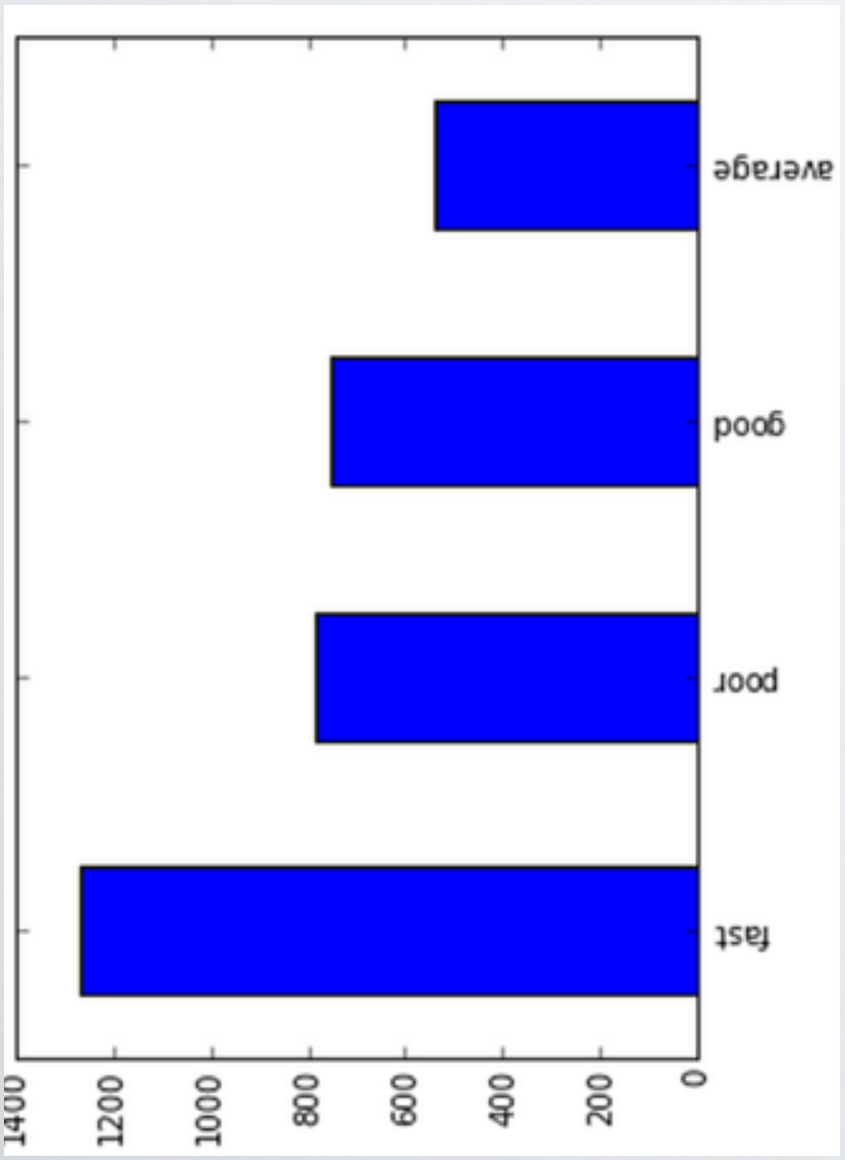
- car_class - SUV, Luxury becomes
  primary_class - SUV
  secondary_class - Luxury

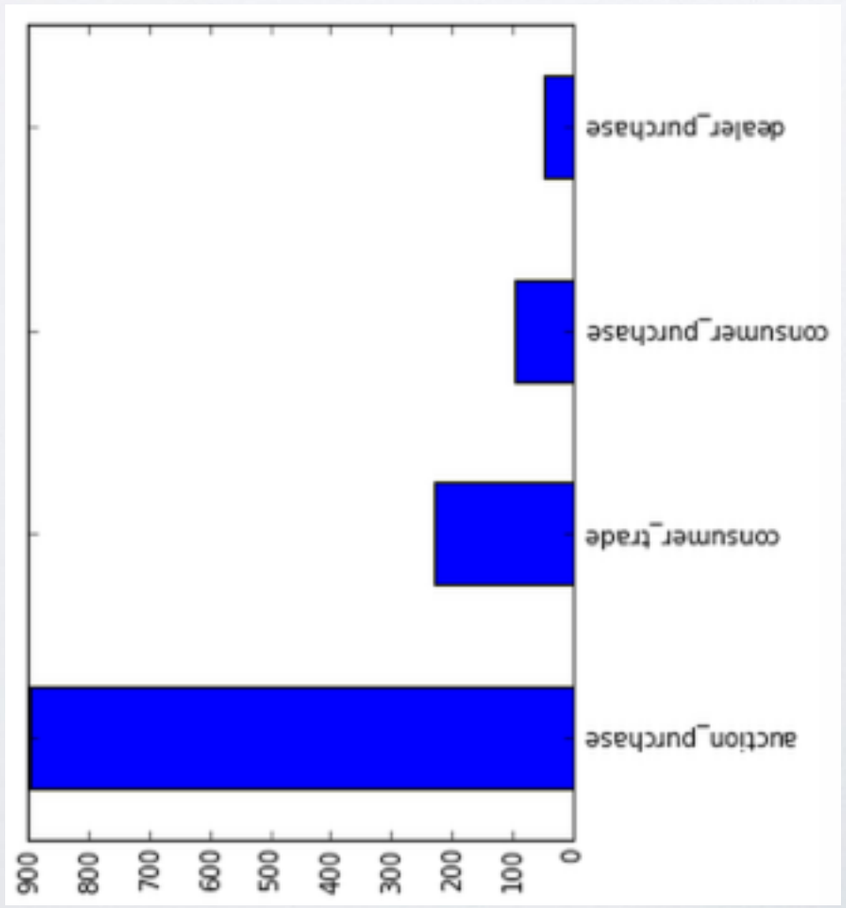





FORD
CHEV
BMW
TOYO
NISS
MB
DODG
PORS
GMC
JEEP
0
50
100
150
200


ebay
autotrader
cargurus
cars.com
known_other
unknown
google
0
100
200
300
400
500
600


fast
poor
good
average
0
200
400
600
800
1000
1200
1400


auction_purchase
consumer_trade
consumer_purchase
dealer_purchase
0
100
200
300
400
500
600
700
800
900

Feature Selection

- Removed age and age group from features since we're predicting these
- Removed features that we know AFTER the car is sold i.e. gross profit, warranty purchases, etc...
- Removed total cost since this is made up of acquisition and reconditioning costs
- Remove features like stock_number and car_class which have more granular features built off of them
- Removed sales_source since this is known after the sale

Feature Selection Contd.

- Ran lasso to eliminate features with zero coefficients
- Looked at feature importance from random forest

	importance
reconditioning_cost	0.107
sale_price	0.098
distance_from_dealer	0.096
acquisition_cost	0.091
mileage	0.084
year	0.031
C(color)[T.Red]	0.010
C(acq_source)[T.consumer_trade]	0.010
C(color)[T.White]	0.010
C(color)[T.Blue]	0.009

Modeling Process

- First ran lasso with unscaled variables and got a mean absolute error of 33.7 days

- Scaled the data and reran lasso to get 0.78 which translates into 38.82 days

- Tried KNN and SVM but they don't support continuous variables

- Tried random forest and got a result of 36.6 days. Did not scale the data since random forest already does this for you

Conclusion/Next Steps

- A mean absolute error of 34 days is ok but ideally want something within 14 days

- Create additional features and try to expand the data set to further reduce the MAE

- Reconditioning cost accounts for 11% of variance within the data set. It'd be interesting to try and predict this as a future project - Has many internal business applications