Predicting Sale Price of Auction Machinery

4/17/2020
Chelsea, Henry, Jaime, Min, Niv

Problem & Data Description

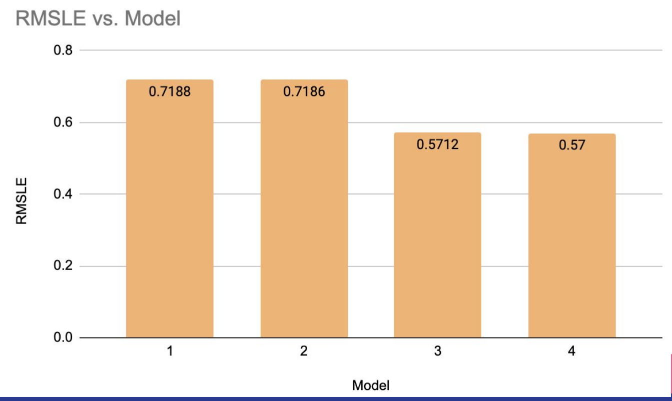
- Chose to predict heavy equipment auction prices
  - Were more interested in prediction over inference
- Training dataset was from one CSV containing > 400,000 entries and > 50 columns
- Data was messy
  - Many of the > 30 machine configuration categorical columns had > 50% null values
- Needed to do feature selection and impute a lot of data
- Wanted to accomplish maximizing model performance on unseen data with tuning through regularization

Team Organization

- Set a goal to understand the data
- Initial skeleton
  - Pull from main repo
- Self-exploration of the data
- Trade notes via slack and Zoom conversation

Process:

- Brainstormed all the features which might impact Sales.
- Created Master helper function: cleaned all the selected features.
- Tested RMSE for Linear_Regression, Ridge and Lasso.
- Baseline Model Score: 0.7188 (SaleYear and YearMade).
- Model-1 Score: 0.7186 (added MachineHoursCurrentMeter).
- Model-2 Score: 0.5712 (added 15 more).
- Model-3 Score:0.570.



Accomplishments

- Starting models
    - Linear
    - Ridge
    - Lasso
- Performance metrics
    - Root Mean Squared Log Error
- Validation
    - We went with a 10 fold validation, with test size of 0.2
- Grid search and found Ridge Regression achieving best results

Performance on Unseen Data

- **Final RMSLE** = **0.5773 on unseen data**
  - Ridge Regression @ alpha 25
  - Model performance on train data: RMSLE = 0.570

- 14 Features Used
  - 9 or 14 features were created from transformations we performed on dataset
  - 4 types of Features used
    - **Date/Age:** YearMade, Saleday, Salemonth, Saledayofyear. Age
    - **IDs:** ModelID, SalesID MachineID
    - **Size**: Tire_Size, ProductSize,
    - **Enclosure:** Enclosure_EROPS, Enclosure_EROPS AC, Enclosure_None Unspecified, Enclosure_OROPS

New Things We Learned

- Important to research & understand features
- How to handle null values by imputing variables
- Cleaning data
- Different model for prediction
- Search for the best model for best score

Appendix

Final Model Feature List

- **Final RMSLE = 0.5773**
  - Utilized Ridge Regression @ alpha 25

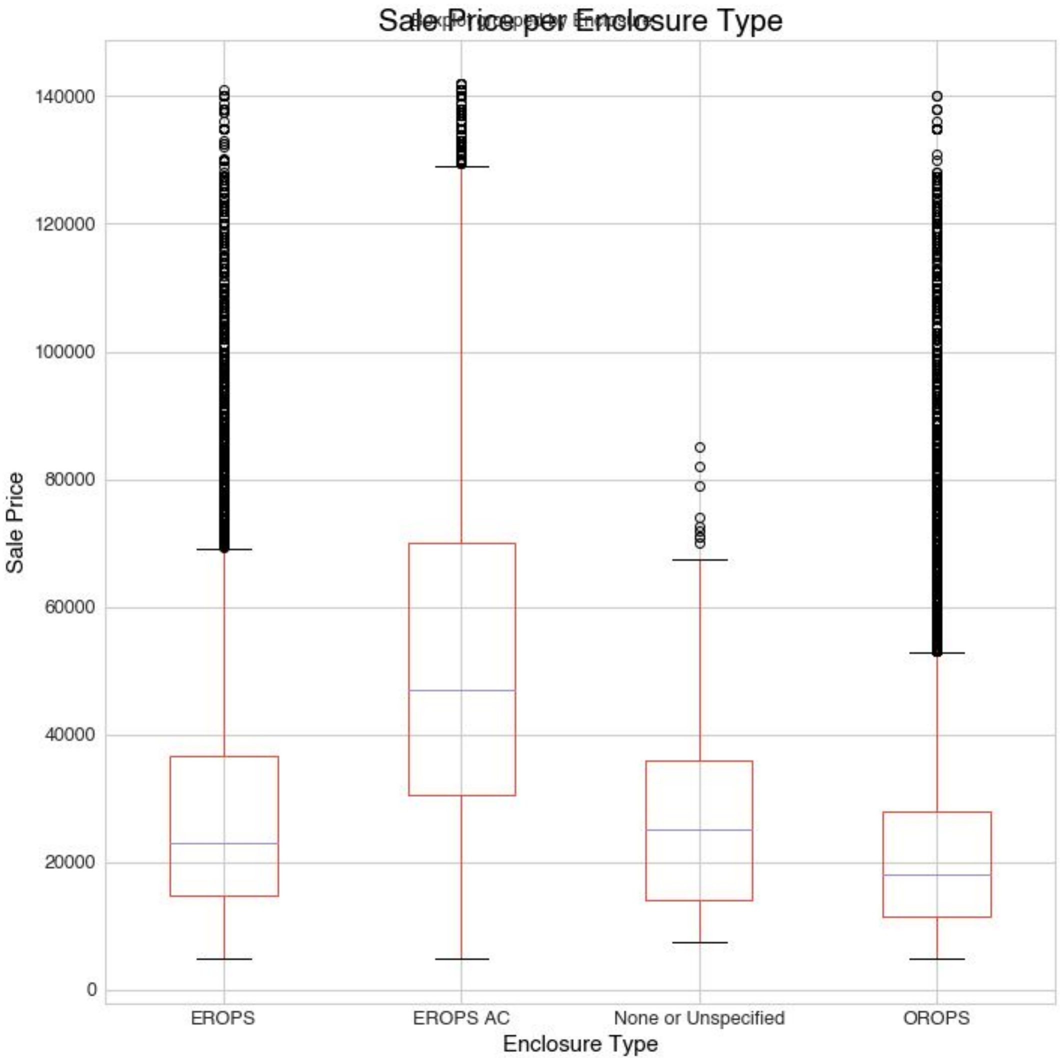
- **Feature List**
  - **Date/Age:** YearMade, Saleday, Salemonth, Saledayofyear. Age
  - **IDs:** ModelID, SalesID MachineID
  - **Size**: Tire_Size, ProductSize,
  - **Enclosure:** Enclosure_EROPS, Enclosure_EROPS AC, Enclosure_None Unspecified, Enclosure_OROPS

Group Work Approach

- One Team Member creates initial branch with:
  - Cursory Cleaning
  - Selection of two three features
  - Brief Transformation
  - Linear Regression run
  - Cross Validation to establish baseline score
    - Set Random Seed so it is consistent
- Split up so each team member can:
  - Do furth EDA
  - Add/Remove features from baseline model
  - Transformations
  - Linear Regression
  - Cross Validation and compare to baseline
  - Repeat
- Take best model, use it on test data, submit!

Brainstorming (delete/move later)
Initial Features:

- MachineHoursCurrentMeter → need to figure out how to handle n
  linear correlation
- YearMade
- Saledate --> transformed to get year only
- UsageBand
- ProductSize
- State
- Drive_System - Chels looked into, may be dead-end because most are unknown
  - `Unique vals: ([nan, 'Four Wheel Drive', 'Two Wheel Driv`
  - `'All Wheel Drive'], dtype=object)`
  - Transformation: Change nan to ‘No’, then map to ints:
  - map({'No':0, 'Four Wheel Drive':4, 'Two Wheel Drive':2, 'All Wheel Drive
- Enclosure - chels looking into
  - `OROPS 173932 - "Open Roll Over Protect`
  - `EROPS 139026 - "Enclosed Roll Over Pro`
  - `EROPS w AC 87820 - treat same as "EROPS"`
  - `EROPS AC 17 - treat same as "EROPS"`
  - `NO ROPS 3 - treat same as "None or`
  - `None or Unspecified 2 - put NaNs into this grou`
  - `Name: Enclosure, dtype: int64`
-


Sale Price per Enclosure Type
Sale Price
0
20000
40000
60000
80000
100000
120000
140000
EROPS
EROPS AC
None or Unspecified
OROPS
Enclosure Type