LEMONAID

Ellen Kim

technical

Background: What is a Lemon car?

Lemon (auto):

In order to qualify as a lemon under most state laws, the car must (1) have a substantial defect covered by the warranty that occurred within a certain period of time or number of miles after you bought the car, and (2) not be fixed after a reasonable number of repair attempts. Varies drastically from state to state.

Basically a lemon car is a dud.

Objective

Create a tool to determine if a used car is likely to be a Lemon or not.

Create recommendation system for potential lemon buyers.

Why?

Business model: attract user traffic by making lemon predictor available.

Users looking at lemons will receive recommendation from our inventory.

1. Consulting Service:
   a. Become “WebMD” of car diagnostics
2. Use lemon predictor as funnel to our inventory. → increase business sales
3. Potential arbitrage opportunities:
   a. prices of cars can vary drastically from state to state.

Data

CARVANA

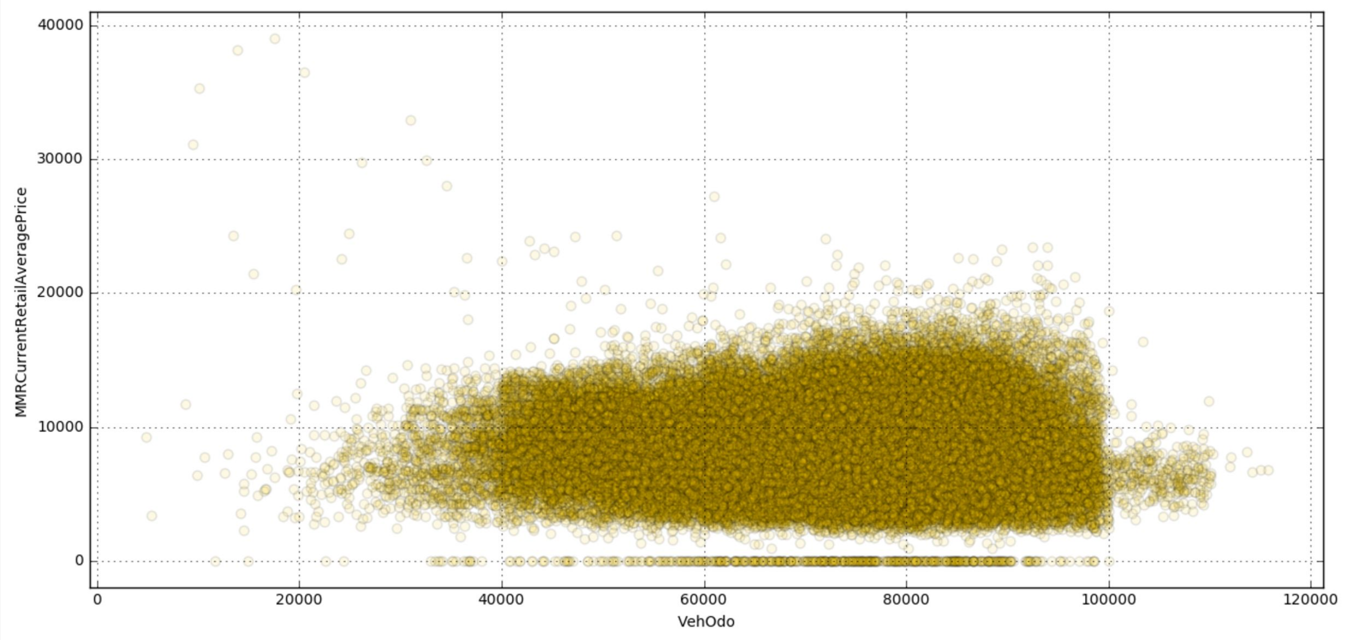
- Dataset (Kaggle competition by CarVana)
- There are 32 Independent variables (C3-C34)
    - Make, model, odometer, color, auction, warranty price, state, etc
- The dependent variable (IsBadBuy) is binary (1 = lemon, 0 = not a lemon).
- The data contains missing values.

EDA

1. Data did not show significant relationships between
    a. Lemon and auction
    b. Lemon and size
    c. Lemon and make
        i. # of Lemons was proportional to output of that make.
    d. Lemon and price:
        i. Lexus and Calidillac's are drastically more expensive when lemons, than not.
        ii. Infiniti, Mini are slightly more expensive when they're lemons than not.

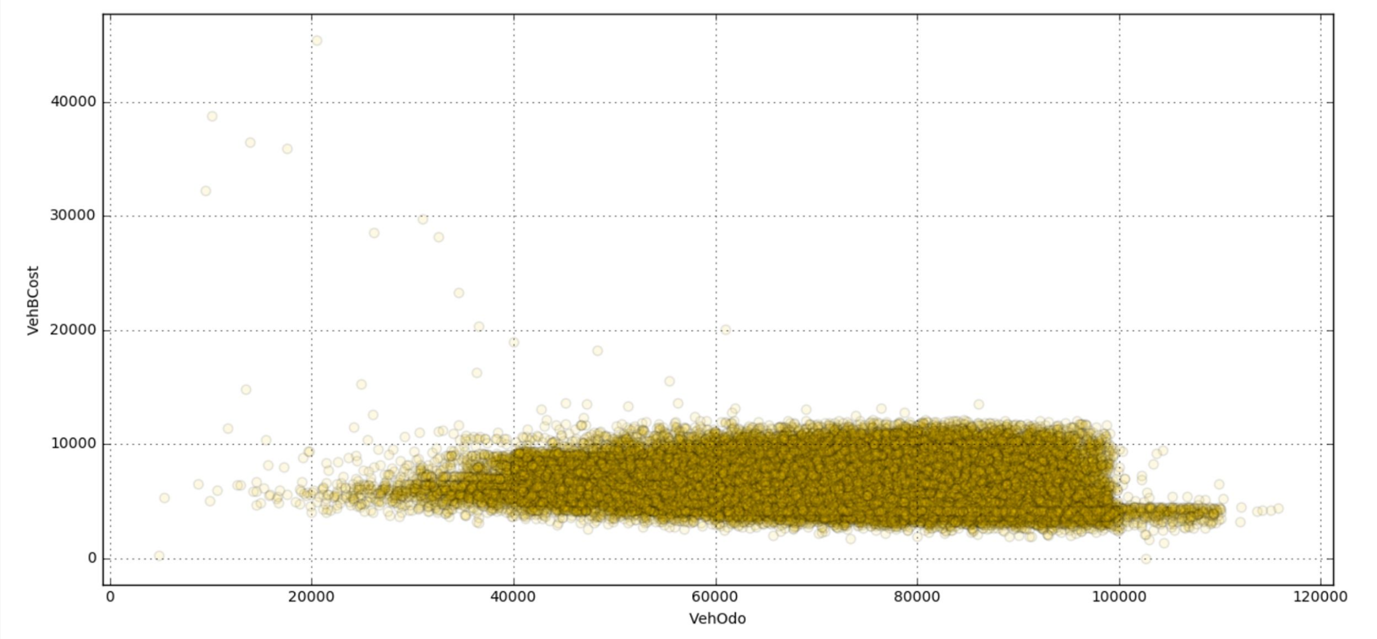
Data: Price

1. Retail Price Range: $3,000 and $19,000
2. Odometer Range: 30k to 100k miles



Data

1. Auctions drastically change price range.
2. Auction price range: $3,000 - $12,000.
3. Clear drop when car reaches 100k miles at $5,00-.



Data Processing

- Missing values:
  - Imputed using KNearestNeighbors for Wheeltype (alloy, covers)
    - Reduced model accuracy by 2%.
  - Dropped Nan values → better performance

- Data Cleaning : normalize, make dummy variables.

Early Algorithm Selection

Ex: Random Forest

```
             pred_lemon  pred_not_lemon
Lemon                 8            1985
Not_Lemon             8           18848


50% (default) THRESHOLD
              precision    recall  f1-score   support

          0       0.90      1.00      0.95     18856
          1       0.50      0.00      0.01      1993

avg / total       0.87      0.90      0.86     20849

The accuracy score for threshold of 50% = 0.90440788527
```

1. Decision Tree
2. Random Forest
3. Logistic Regression
4. AdaBoost
5. Gradient Boost

- Overfitting caused model to label all cars as “Lemons”.
- Could not predict likelihood of Lemon.
- Baseline is ~90%.

Other attempts to increase model performance

1. Balancing data
   a. Only ~10% of data were lemons.
2. Fine tuning hyper-parameters
   a. Penalty
   b. Inverse of regularization strength (C)

Model of choice

1. Logistic Regression

Confusion Matrix for 50% threshold

	pred_lemon	pred_not_lemon
Lemon	2	1991
Not_Lemon	6	18850

50% THRESHOLD

	precision	recall	f1-score	support
0	0.90	1.00	0.95	18856
1	0.25	0.00	0.00	1993
avg / total	0.84	0.90	0.86	20849

Fine-Tuning

- Best Model:
  - Logistic Regression w/ 12% threshold.
- Accuracy = 73.92%
- Baseline = ~90% likely to be non_lemon
- Accuracy improvement is not the only way to add value to the business.
- Goal: increase precision

Confusion Matrix for 12% threshold

	pred_lemon	pred_not_lemon
Lemon	1016	977
Not_Lemon	4461	14395

12% THRESHOLD

	precision	recall	f1-score	support
0	0.94	0.76	0.84	18856
1	0.19	0.51	0.27	1993
avg / total	0.86	0.74	0.79	20849

Recommendation System

Used Nearest Neighbors based on following features:
Generally , non-negotiables for buyers.

1. Make and model
2. Price
3. Mileage
4. Size

Rec System: Feature Selection

Would you buy a gold car?

Excluded features:

1. Color
2. State
   a. Price varies drastically by state.
   b. Could find financial savings by shipping car than purchasing locally.
3. Year of manufacturing

Rec. System: Process

1. Split Data:
   a. X_test → hypothetical inventory
2. Logistic Regression
   a. Kept only non-lemons
3. Recommended cars from inventory of non-lemons

Flask

Miles on Odometer 100000

Price of Vehicle 10000

Warranty Cost 1200

Auction Adesa

Month of Purchase January

Year Car Manufactured 2001

Vehicle's color RED

Make-Model CHEVROLET MALIBU

State NJ

Nationality of Vehicle AMERICAN

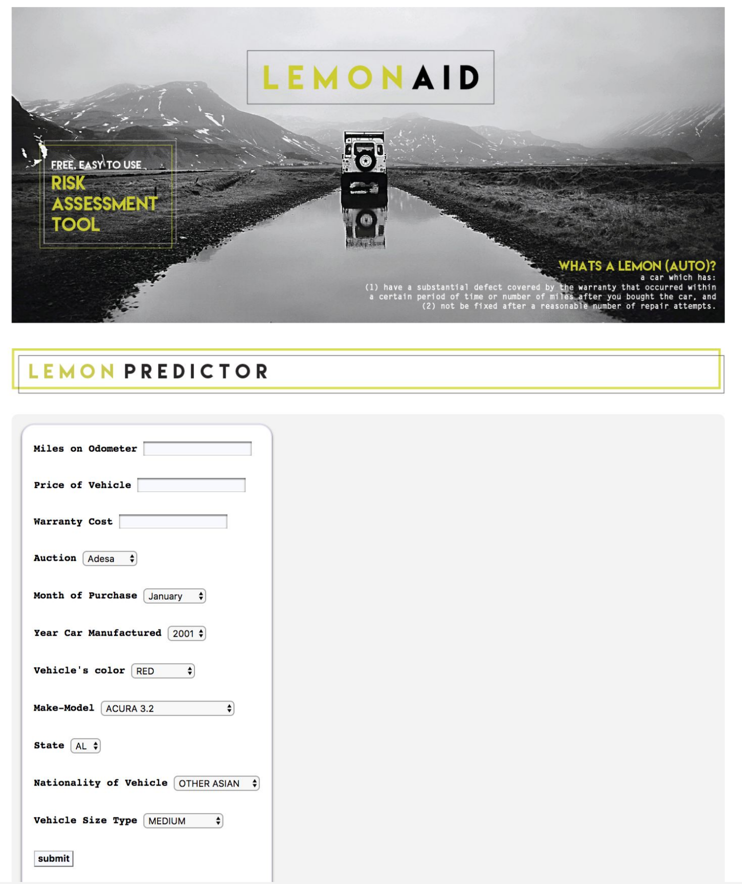
Vehicle Size Type COMPACT

submit

localhost:5000/result

Apps Bookmarks Kindle Cloud Reader HauteLook Gmail Itin Repo


```
{
  "The vehicle you have entered is": "likely to be a lemon",
  "rec_car1": [
    "Odometer:98581.0",
    "Retail Price: 9074.0",
    "MANHEIM",
    "10",
    2006,
    "GREY",
    "AMERICAN",
    "CHEVROLET MALIBU",
    "MEDIUM",
    "AZ"
  ],
  "rec_car2": [
    "Odometer:96922.0",
    "Retail Price: 9752.0",
    "MANHEIM",
    "9",
    2007,
    "WHITE",
    "AMERICAN",
    "FORD FOCUS",
    "COMPACT",
    "TX"
  ],
  "rec_car3": [
    "Odometer:98726.0",
    "Retail Price: 8926.0",
    "MANHEIM",
    "10",
    2006,
    "RED",
    "AMERICAN",
    "CHEVROLET MALIBU",
    "MEDIUM",
    "CA"
  ],
  "rec_car4": [
    "Odometer:96454.0",
    "Retail Price: 9570.0",
    "MANHEIM",
    "12",
    2006,
    "SILVER",
    "AMERICAN",
    "CHEVROLET MALIBU",
    "MEDIUM",
    "AZ"
  ],
  "rec_car5": [
    "Odometer:97765.0",
    "Retail Price: 8884.0",
    "OTHER",
```

Conclusion

1. Can’t predict if a car will be a lemon based on these features.
    a. These may not be the determining features.

What we need

1. A decision making tool that takes more comprehensive approach:
    a. Monthly installments
    b. Financing interest
    c. Cost of insurance
    d. Average cost of gas
    e. With risk assessment
2. Incorporate CarFax information
    a. Model must include condition of individual vehicles (regular maintenance, accident history)

Questions?