Stats141 Final Project

Jiahao Huo
Minjie Xia

Project Overview

- Datasets from DataFest 2015 by *Edmunds.com*
    - Visitors
    - Leads
    - Configuration
    - Transactions
    - Shopping
- *Edmunds.com*: an American online resource for automotive information

“if a customer leaves information on the website for a particular car, is he/she going to buy the car?”

Project Goal

- Answer the question
    - Building a binary logistic model
- Aim
    - Determine how likely a certain customer will buy cars
    - Then, decide if to pursue this customer or not

Data Cleaning and Feature Engineering

- Create Response Variable
  - 1: customers who both left contact information and bought cars
  - 0: customers who only left contact information but did not buy cars
- Selected Key Features from Leads dataset
  - “dealer_distance”, “model_year”, “make”, “model”, and “style”
  - the last four are transformed to be binary to indicate if there is a value for each observation

Data Cleaning and Feature Engineering

- Engineered 12 new variables from Leads, Shopping, Configuration and Transactions datasets

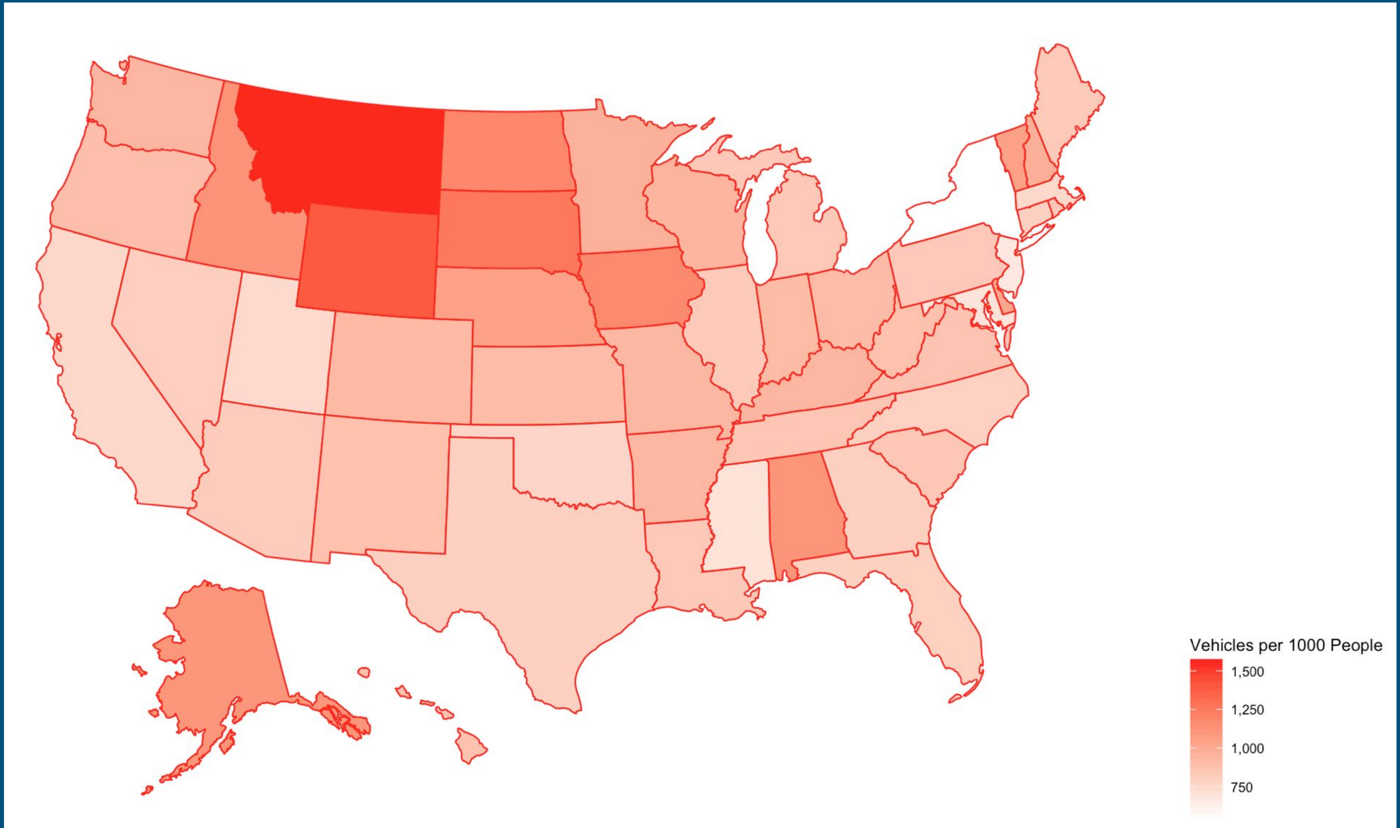
Name	Type	Description
contactinfo_n	integer	number of times a certain customer left information on the website
shoppingdate_n	integer	how many days a customer viewed the website
diffcar_n	integer	how many different cars a customer viewed

Data Cleaning and Feature Engineering

Name	Type	Description
bcarview_n	integer	the number of times a customer viewed a car which he/she eventually bought
bmakeview_n	integer	the number of time a customer viewed a make which he/she eventually bought
singleday_max	integer	the maximum number of webpages a customer has gone through on a single day
year_diff	integer	the difference in terms of year between the year when customers left their contact information and the year of the model they were interested in

Data Cleaning and Feature Engineering

Name	Type	Description
Impinfo_n & lessinfo_n	integer	how specific a customer’s requirements for a car were when leaving information: model year, make, model, style, body type, trim, interior color, exterior color, interior fabric color, fuel, engine, and transmission.
new, old, & cpo	binary	new or old or certified pre-owned cars
ppfY & ppfN	binary	price promise flag or not
“top10”, “top20”, “top30”, “top40”, and “top50”	binary	With external data: the state ranking of vehicle per 1000 people



Map of Each US State’s Vehicles per 1000 People

Data Cleaning and Feature Engineering

- Eliminated features that has more than 80% missing values
- The remaining features:
    - "year_diff", "leads_month", "impinfo_n", "lessinfo_n", "new", "old", "cpo", "ppfY", "ppfN", "top10", "top20", "top30", "top40", and "top50"

Visitor Data

- Group 1: Features that have great influence on the response variable
  - Example: Flag to identify if a visitor ever viewed new vehicle pages
- Group 2: Features that have been turned into binary
  - Example: Page view count for dealer reviews index, page view count for long-term road tests
- Group 3: Categorical features
  - Example: Credit levels, age groups
- Group 4: Continuous features
  - Example: Time in seconds spent on New vehicle pages, total count of distinct models that a visitor key has viewed

visitor_key	credit_worthiness
5050115223919190000	Very Good
-2934019458926960000	Excellent
9194032848870020000	Excellent
-3015230026366520000	Good
-5862269435718710000	Very Good
-3959660787836830000	Fair
-4302020546031280000	NA
6301958214233150000	Poor

credit_VeryGood	credit_Excellent	credit_Good	credit_Fair	credit_Poor
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	0
0	0	0	0	1

```
> quantile(dat1$new_page_views,probs = c(0,0.33,0.66,1))
  0%  33%  66% 100%
   1   21   59  486
```

visitor_key	new_page_views
-4303898956343120000	169
-2337460483917090000	31
1344493155451610000	13
-5496533844886550000	6
7348157318923610000	9
8198743122555140000	16
2964366380892430000	110

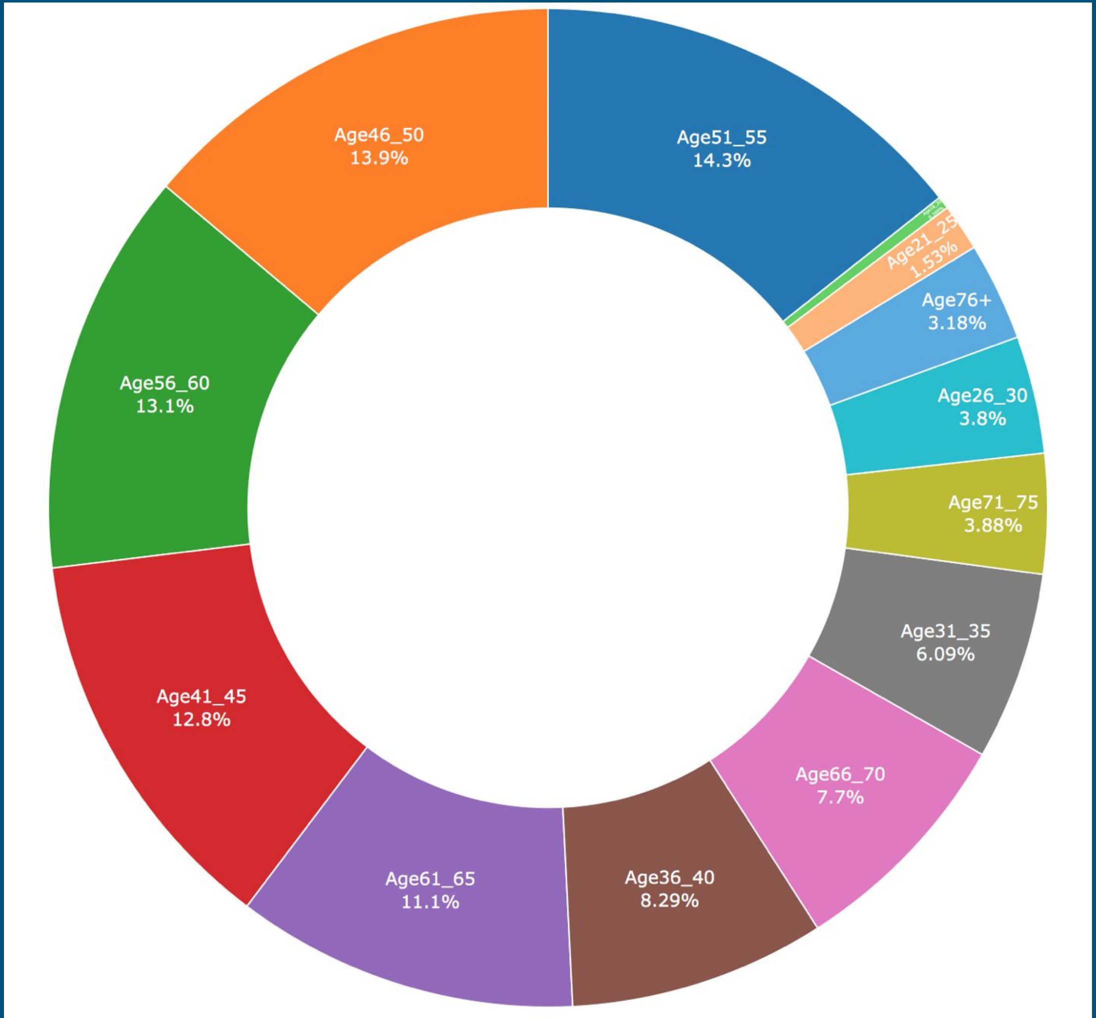
new_page_views_L	new_page_views_M	new_page_views_H
0	0	1
0	1	0
1	0	0
1	0	0
1	0	0
1	0	0
0	0	1

Modeling

- Split data into training (25%) and validation (75%)
- Use Chi-square statistic to filter features for dimension reduction (137 to 47)
- Standardize all the numeric features
- Activate parallel computing capabilities
- Hyper-parameter tuning using optimization methods with random search
    - Algorithm: XGBoost
    - Parameters:
        - eta: shrinks the feature weights to avoid overfitting
        - nrounds: number of rounds for boosting
        - max_depth: the maximum step we allow each tree’s weight estimation to be
        - Evel_metric = “auc”
- 3-fold cross validation to measure improvements
- After 100 times of tuning, the optimal hyper-parameters with the highest AUC value (eta = 0.0608, max_depth = 9, nrounds = 376)

Age Group Distribution

Among all the leads observations which actually bought a vehicle, the distribution of age groups indicates that a customer aged between 40 to 65 has a high probability to purchase a vehicle.



Analysis and Interpretation

- Test optimal hyper-parameters on the testing data
- With binary logistic model:
  - 1: probabilities larger than 0.5 (customers that will buy cars)
  - 0: probabilities lower than 0.5 (customers who will not buy)
- Confusion Matrix

Actual/ Predicted	**0**	**1**
0	1579183	241
1	239916	157

Analysis and Interpretation

- Confusion Matrix
  - Accuracy: 86.8%
  - Misclassification rate: less that 13%
  - False Positive rate: 0.015%
  - High specificity rate
- True Positive rate:
  - low prevalence in the data: only 13% of people bought cars.
  - This low percentage of buying population in the sample may skew the model
  - a high specificity rate and a low true positive rate

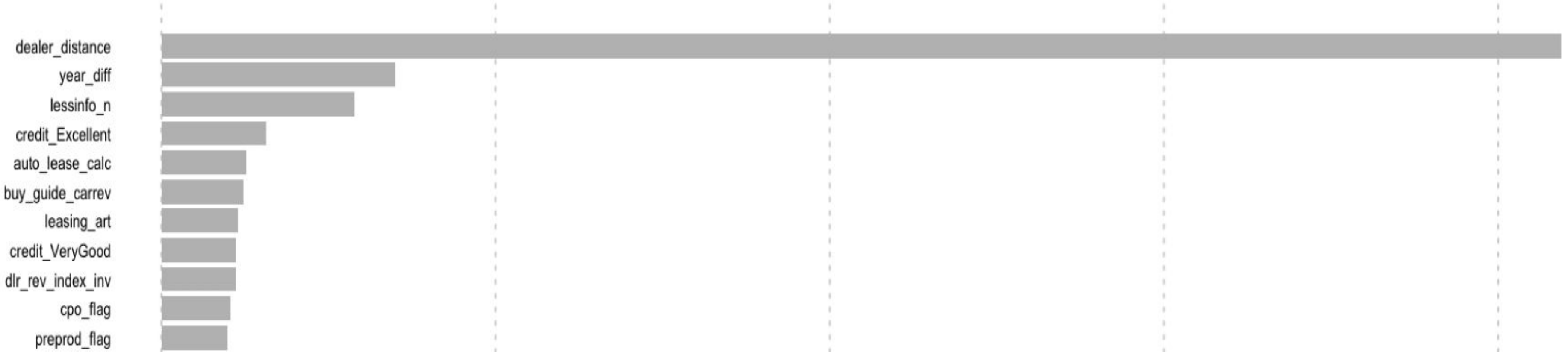
Feature Importance

Gain is the improvement in accurate brought by a feature to the branches it is on.

Feature	Gain	Cover	Frequency	Importance
dealer_distance	0.419	0.560	0.469	0.419
year_diff	0.070	0.075	0.087	0.070
lessinfo_n	0.058	0.033	0.068	0.058
credit_Excellent	0.031	0.017	0.027	0.031
auto_lease_calc	0.025	0.018	0.014	0.025
buy_guide_carrev	0.024	0.007	0.021	0.024

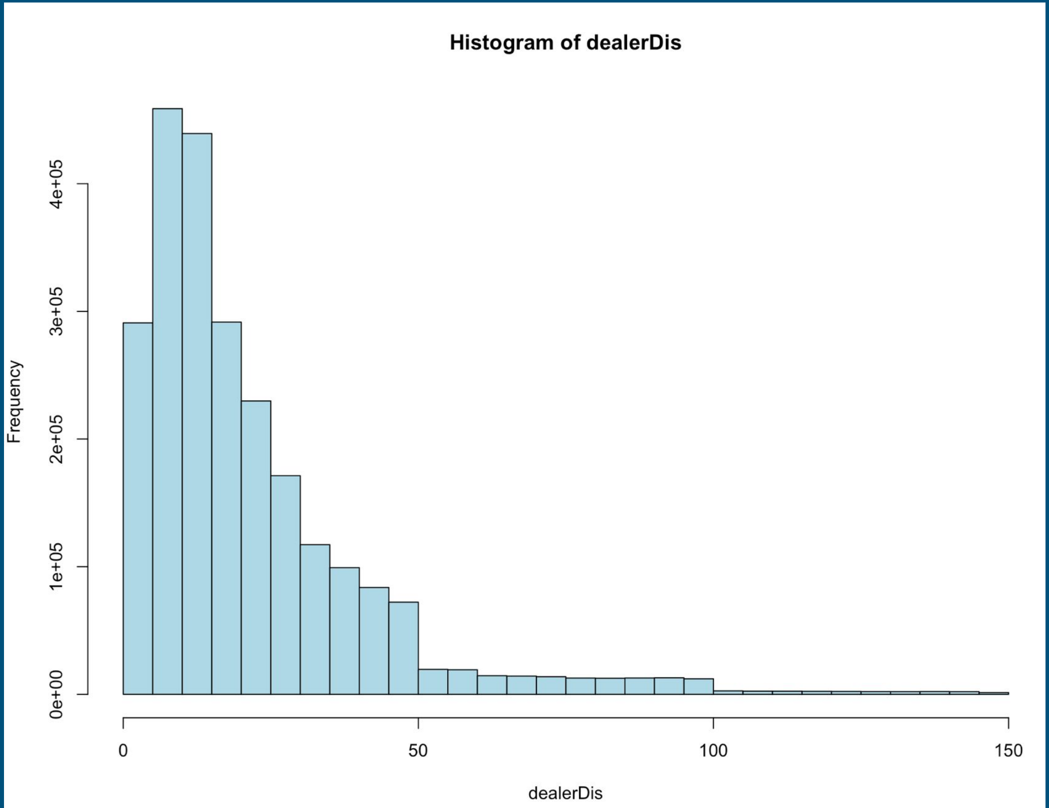
Feature Importance Visualization

- Top 1: dealer_distance
- Top 2: year_diff
- Top 3: lessinfo_n



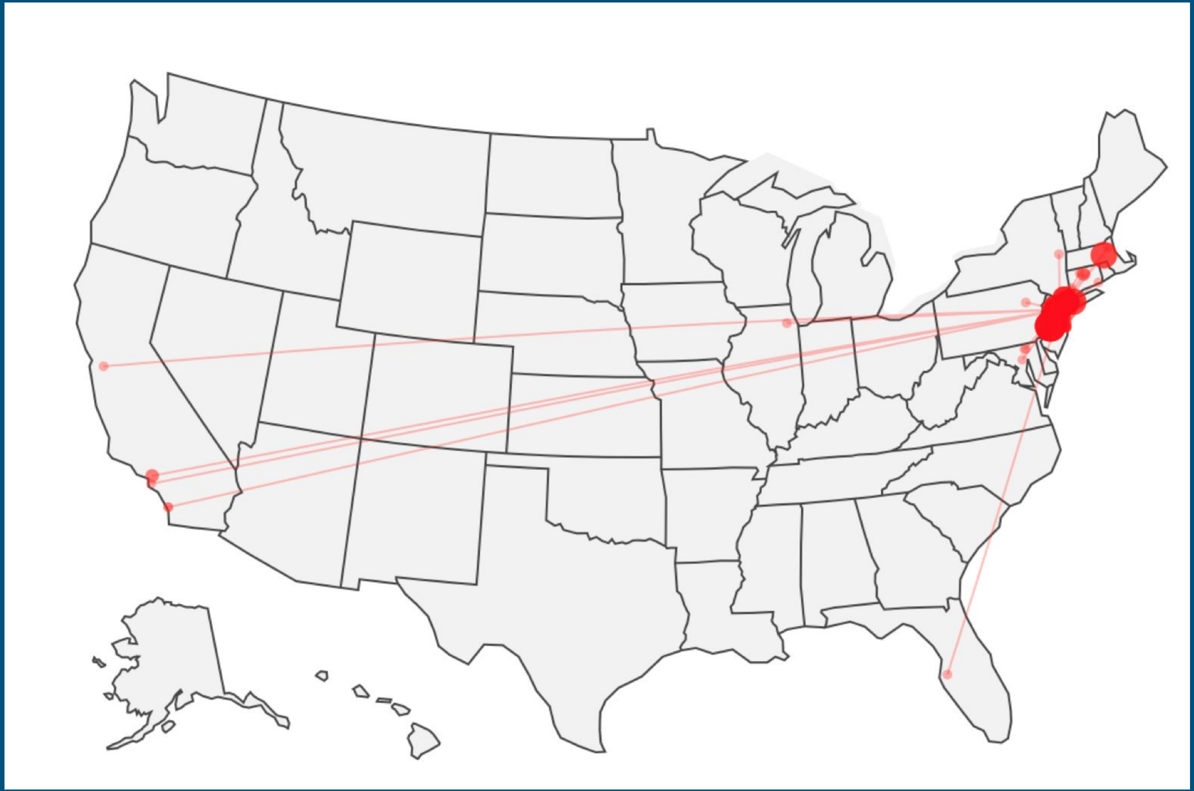
Histogram of dealer_distance

- Highly-skewed (93.1% of the customers who leave their information live within 50 miles from their interested dealer stores)



Map example of dealer_distance

An example showing the dealer distance from one dealer to its customers



Conclusion

- Pertinent question answered:
  - if a customer leaves information on website for a particular car, is he/she going to buy the car?
- Engineered new 113 features and selected 24 variables from original datasets
- 137 features for machine learning
- 47 features for final modeling

Conclusion

- Binary logistic regression with XGBoost
- Hyper-parameter tuned
  - Eta
  - Nrounds
  - Max-depth
- Accuracy of the model: 86.8%

Thank you for your attention