Final Project: STAT8051

Jeonghwan Lee, Jinwen Fu, Jooyong Lee, Seungwon Lee

University of Minnesota Twin-Cities

December 12, 2022

Outline

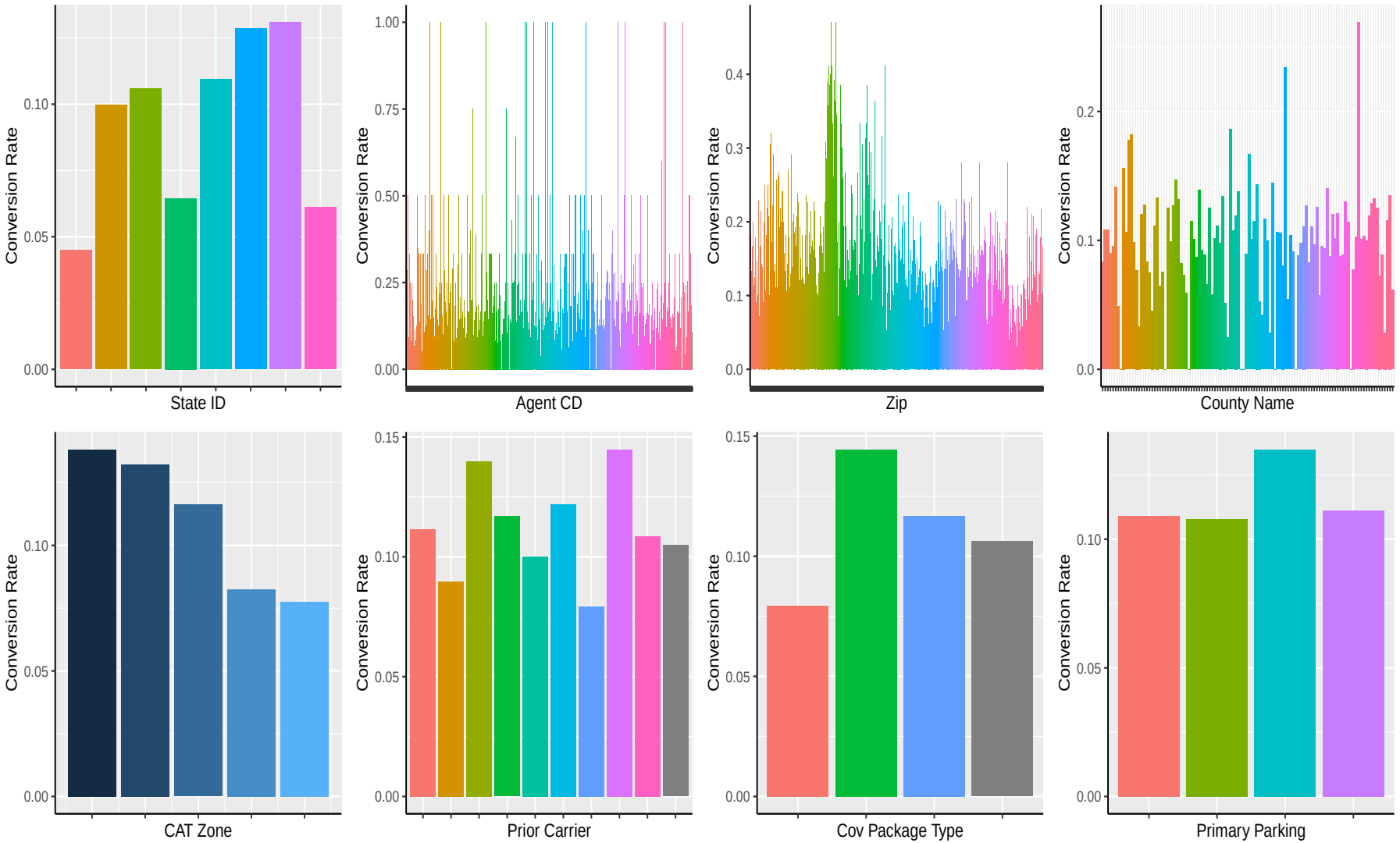
Data Overview

	Policy	**Driver**	**Vehicle**
Observation	49162	106294	169237
Number of Variables	18	5	5
NA(Missing value)	24625	670	1608

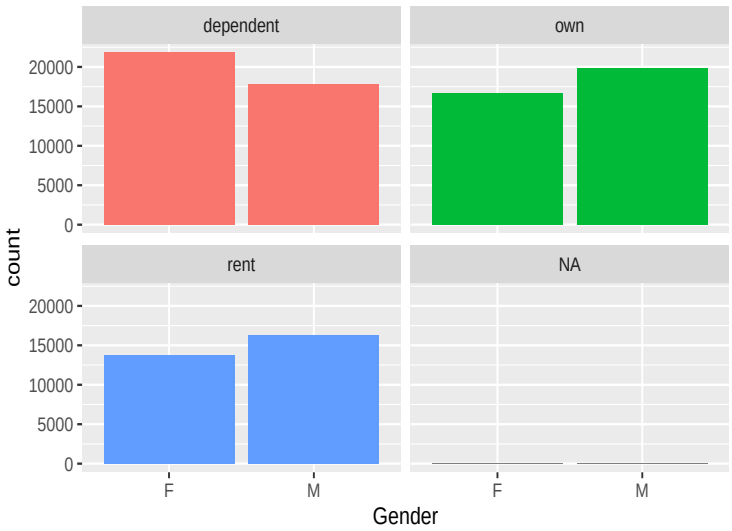
Table: Data overview

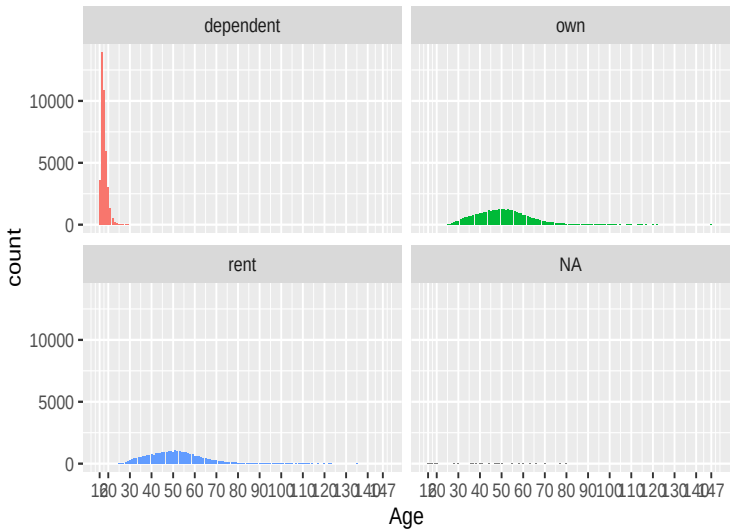
- The data is consist of data sets(**Policy**, **Driver**, **Vehicle**).
- **Policy** is consist of 49162 unique *policy-id*.
- **Driver** and **Vehicle** contain multiple observations for each *policy-id*.

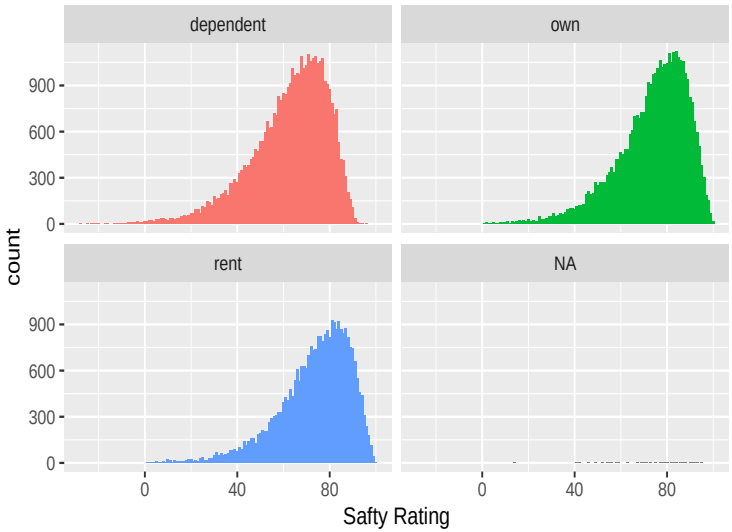
Data Overview-**Policy**

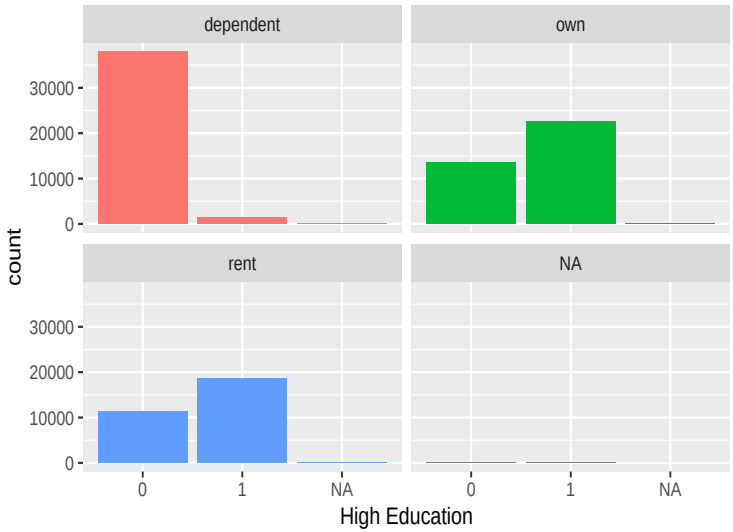


Data Overview-**Driver**

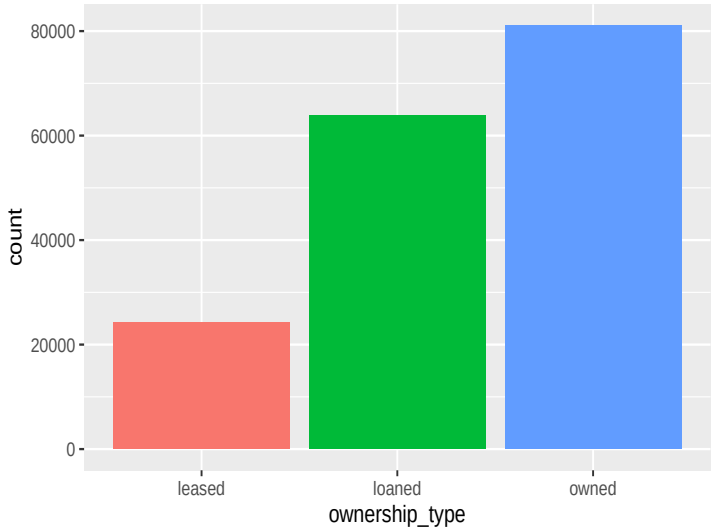


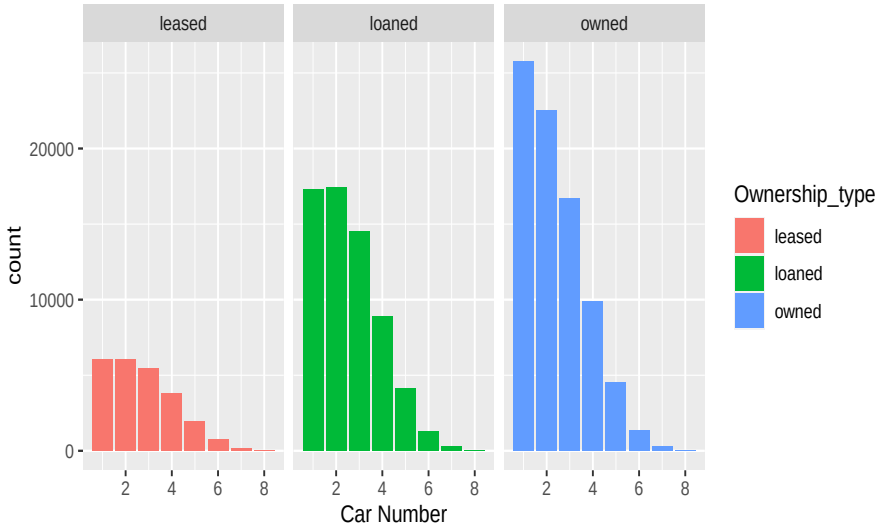


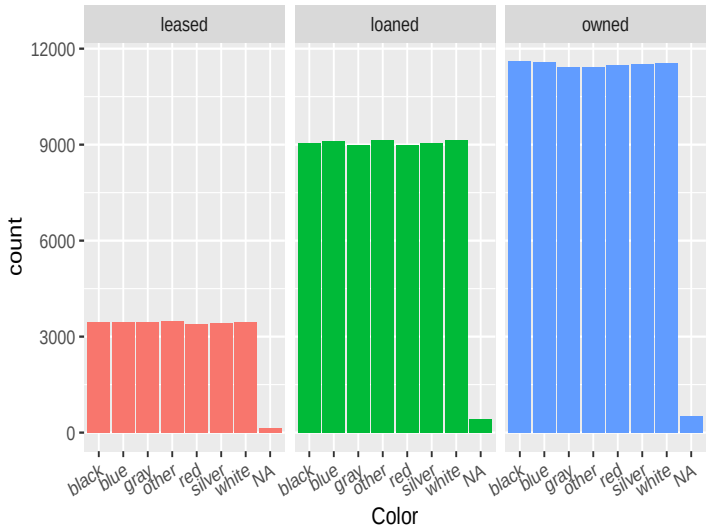


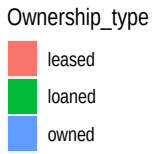


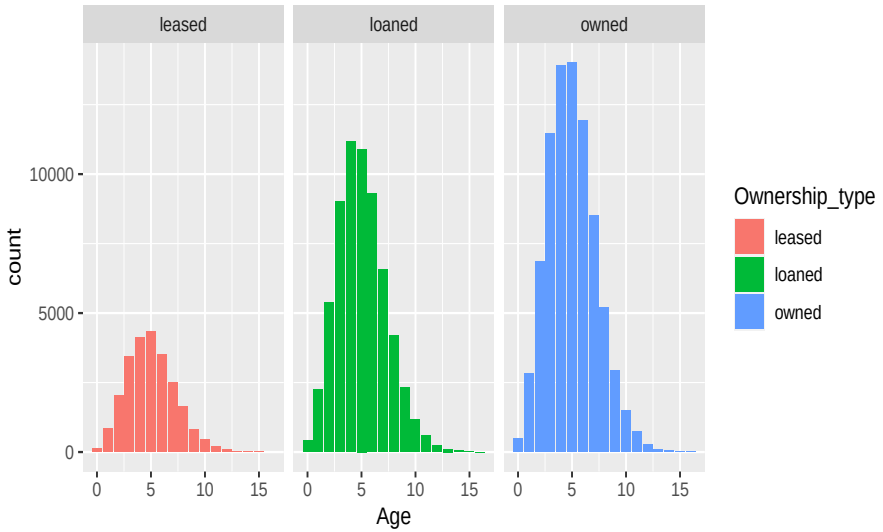
Data Overview-**Vehicle**











Data Overview-Summary

- From the **Policy**, there are apparent differences in *Conversion Rate*, among different groups (*Conversion Rate* is proportion of *convert-ind=1* for each group).
- From the **Driver**, distribution of variables is different according to *Living Status*.
- From the **Vehicle**, there are no apparent differences in the distribution among variables.

Data Pre-processing

- **Goal 1**: Fill missing values with proper values for all data sets.
- **Goal 2**: Convert *zip* to *longitude* and *latitude*.
- **Goal 3**: Merge information from **Vehicle** and **Driver** to **Policy**.

Handling Missing Values

Predictor	**# of NA**
zip	472
Agent_cd	5430
quoted_amt	112
Prior_carrier_grp	5000
Cov_package_type	770
CAT_zone	250
n_safty_rating	77

1. Delete rows with NA
2. Naive Imputation: Filling NA values with the mean(mode) of each columns.
3. **Multiple Imputation**: Imputing NA values using information of non-missing values.

zip to *lon*,*lat*

state	*county*	*zip*	*lon*	*lat*
NY	Bronx	10465	-73.82426	40.82622
FL	Miami-Dade	33141	-80.1484407	25.8535635
WI	Milwaukee	53210	-87.9751299	43.0671805

Table: Example of converting *zip*

- By using Google API **geocode**, we could convert *zip* into *lon* and *lat* data.

Merging Predictors

- The **Driver** and **Vehicle** data was merged into **Policy** data by calculating the mean(mode) of each predictors.

policy_id	*safty_rating*	*age*	*living_status*
policy_5	74	60	rent
policy_5	30	20	dependent
policy_5	29	16	dependent

Table: Driver

policy_id	*safty_rating*	*age*	*living_status*
policy_5	44.333	32	dependent

Table: Policy

Models

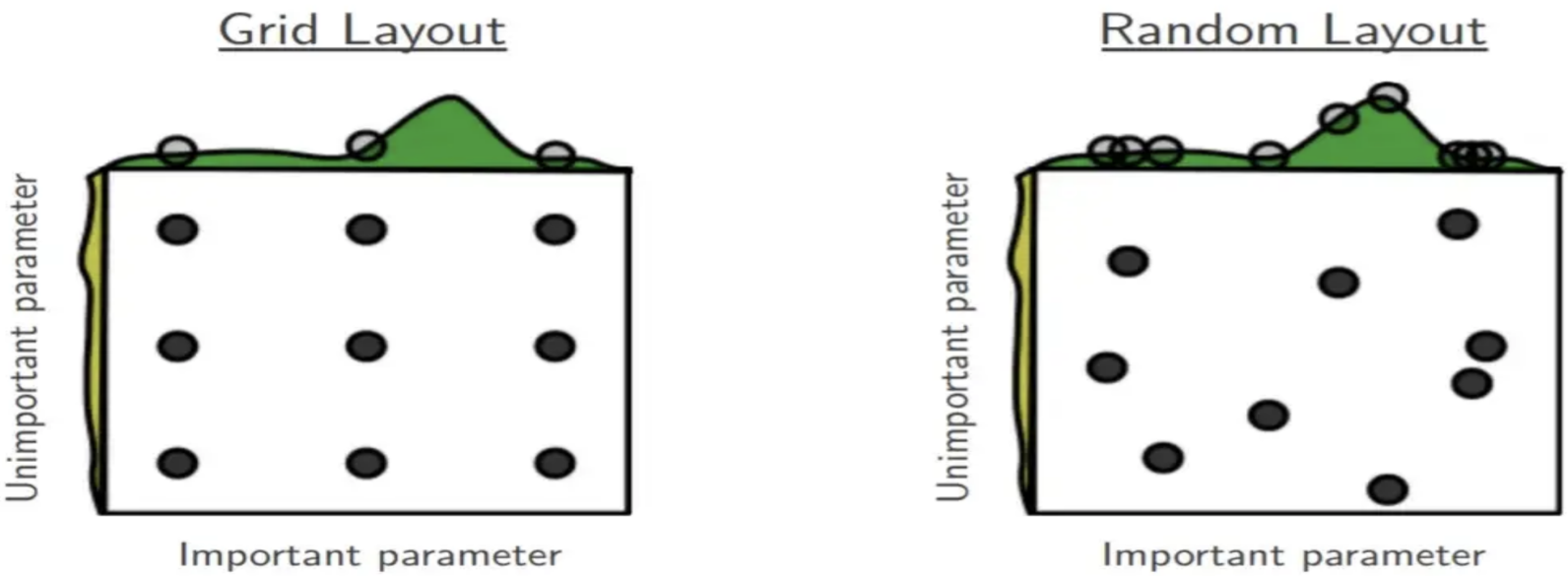
models	best in-sample AUC	out of AUC
Logistic Regression	0.6382	-
Randomforest	0.6597	-
XGBoost	0.6838	0.67665

Table: Model Performances

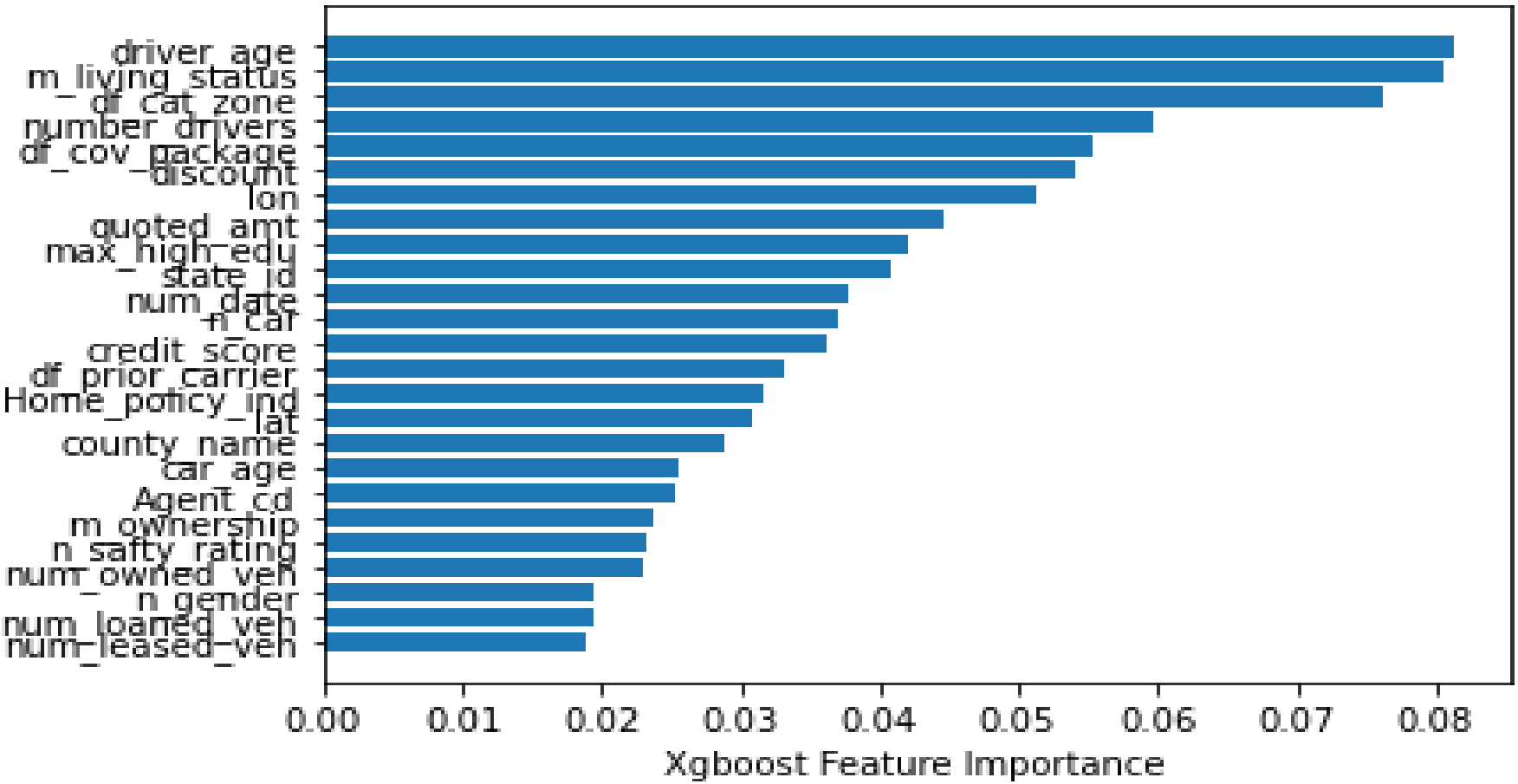
- XGBoost shows the best in-sample AUC, and we use it as a model for the test set.

Hyperparameter Tuning

- RandomizedSearchCV: Optimization of accuracy through one or few "random" parameters. We chose it as this one can outperform the search grid when only a few parameters affect the outcome of our models.



Experiments-Feature Importance



What the Model Suggests

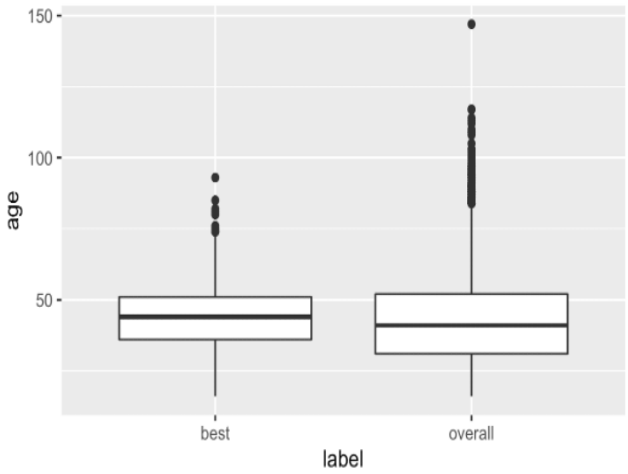
- Use Logistic regression to illustrate the black box
- On training set: prediction $\sim$ important features

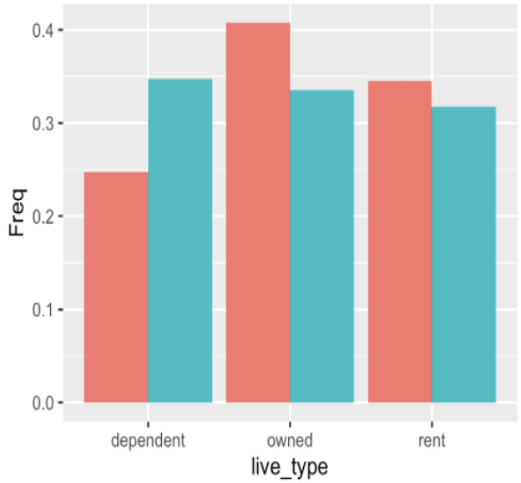
predictor	*coefficient*	*significance*
driver age	2.154e-03	***
living_dependent	0	***
living_own	1.277e-01	***
living_rent	1.271e-01	***
cat zone	-2.232e-02	***
driver number	-2.337e-02	***
discount	4.483e-02	***
quote amount	-3.912e-06	***

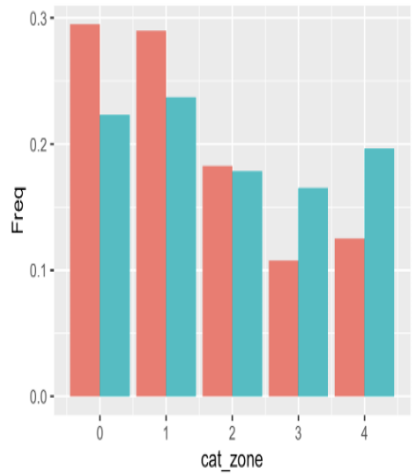
Table: Feature Coefficients

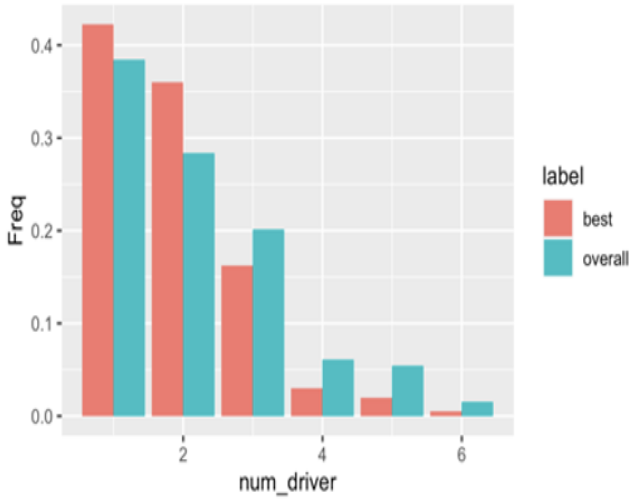
User Persona

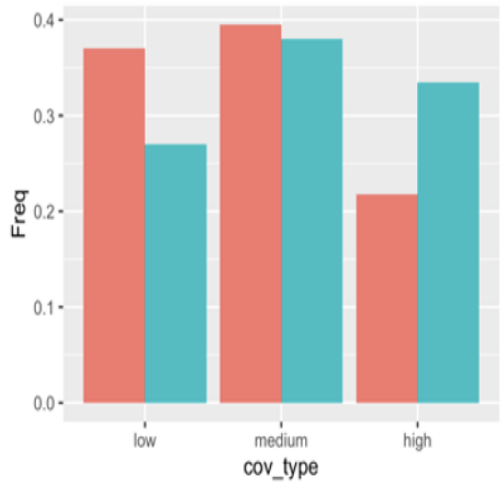
- Select the policies whose predicted probability is near 1

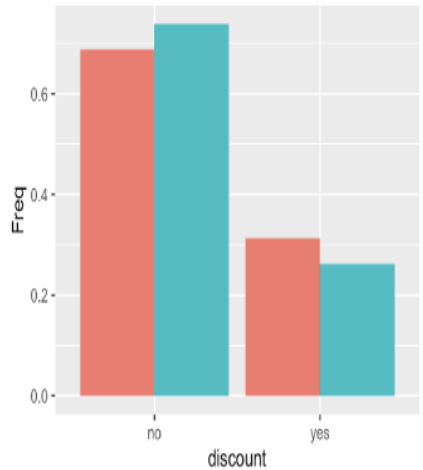












User Persona

- How does a most likely Costumer look like?

predictor	*value*
driver age sum	around 44
driver number	1 or 2
living status	rent or owned
cat zone	1
prior cov package	low
discount	yes
quoted amount	around 4324.75
high education	yes
state id	FL, NJ or NY

Table: User Persona