Conversion Classification Model and Recommendation

Xiaoyun(Clara) Wang ; Yuting Xu

Umass Amherst

Dec 2 ,2019

Outline

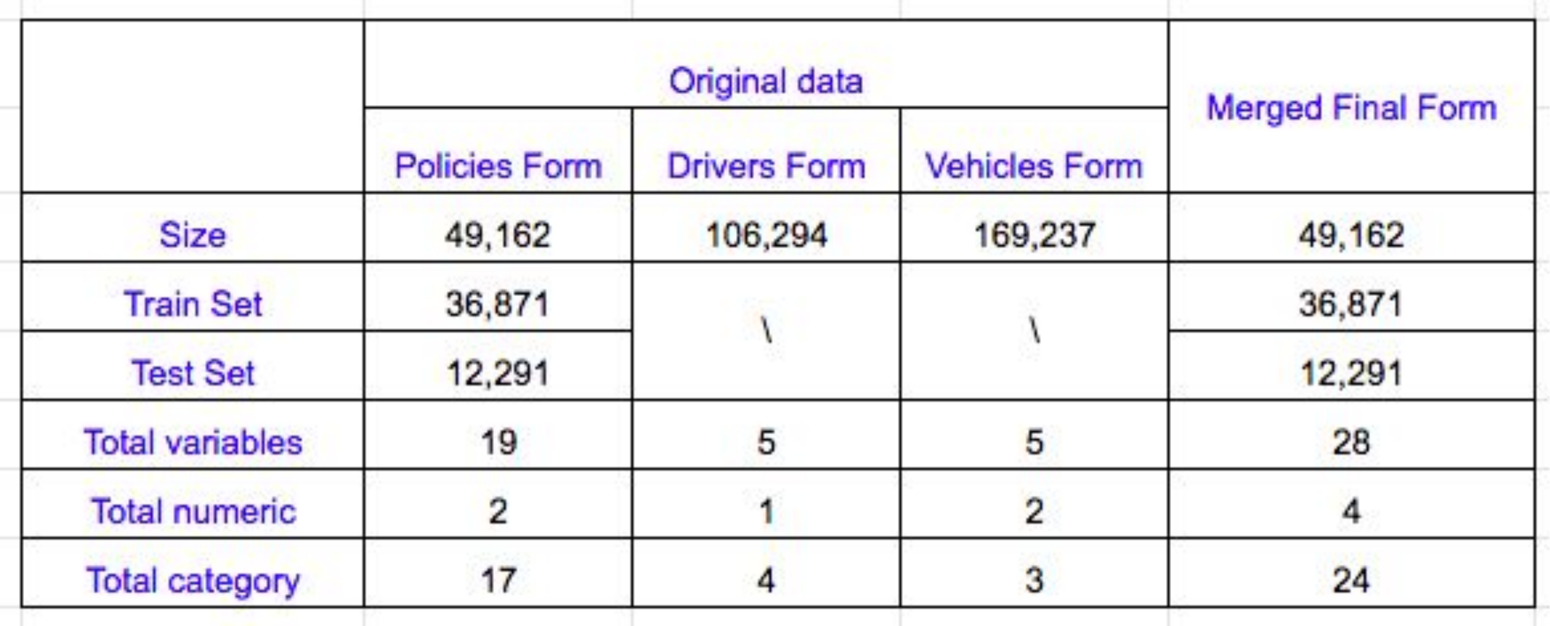
❏ Data Preprocessing

❏ Exploratory Data Analysis

❏ Predictive Modeling

❏ Recommendation

Data Preprocessing - Summary



	Original data			Merged Final Form
	Policies Form	Drivers Form	Vehicles Form	
Size	49,162	106,294	169,237	49,162
Train Set	36,871	\	\	36,871
Test Set	12,291			12,291
Total variables	19	5	5	28
Total numeric	2	1	2	4
Total category	17	4	3	24

Data Preprocessing - New Variables

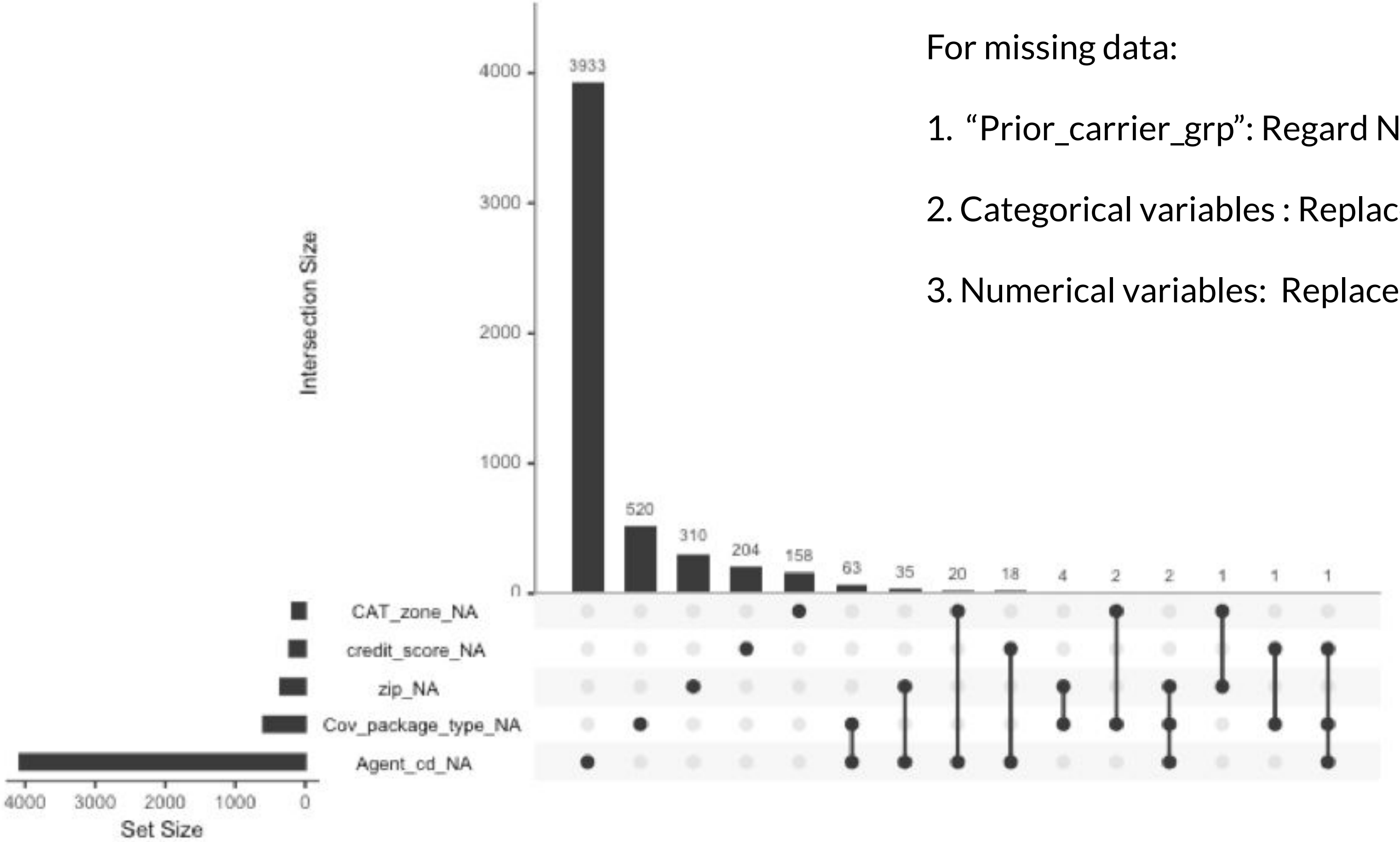
- From drivers dataset

  1. Avg_age_driver (numerical)
  2. High_education_ind (0,1)
  3. Living_status (own, rent, other)

- From vehicles dataset

  1. Avg_age_vehicle (numerical)
  2. Luxury_motor (0,1)
  3. num_luxury_motor(0-8)

Data Preprocessing - Missing Value



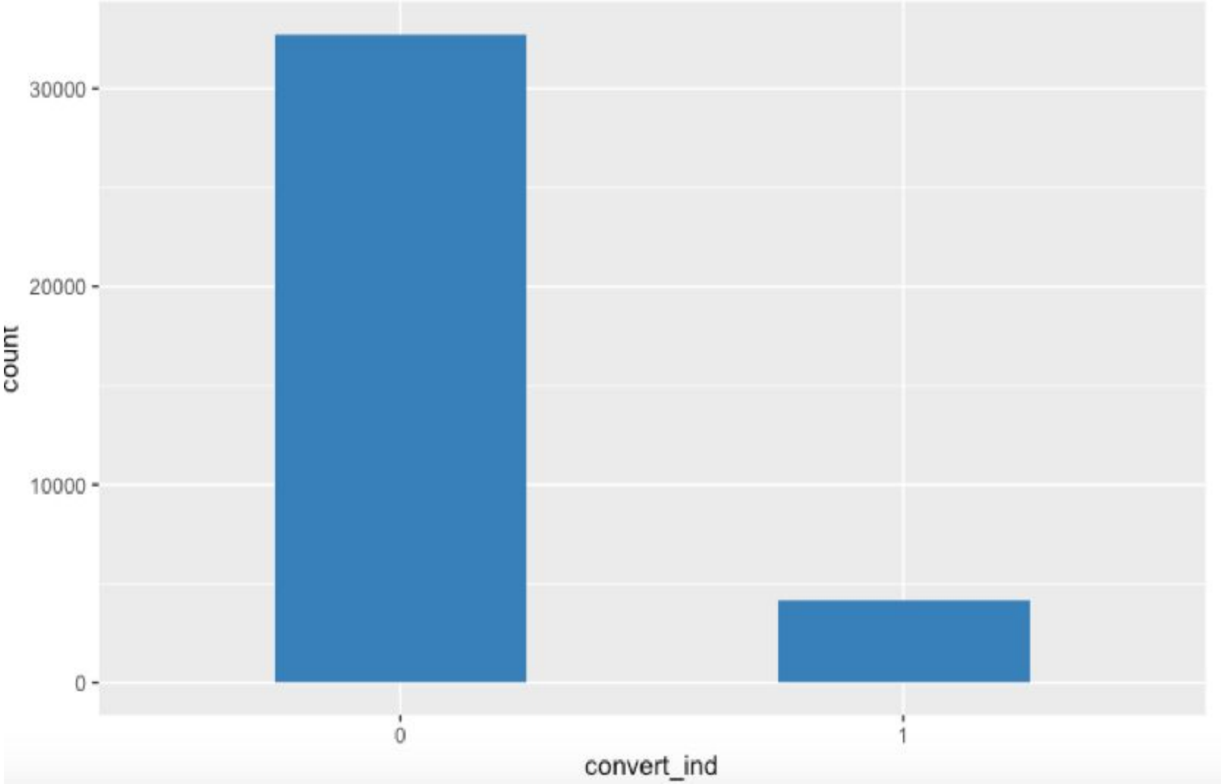
For missing data:

1. “Prior_carrier_grp”: Regard NA in as a new level.

2. Categorical variables : Replace NA in with mod.

3. Numerical variables: Replace NA in with mean.

Exploratory Data Analysis- Convert Indicator



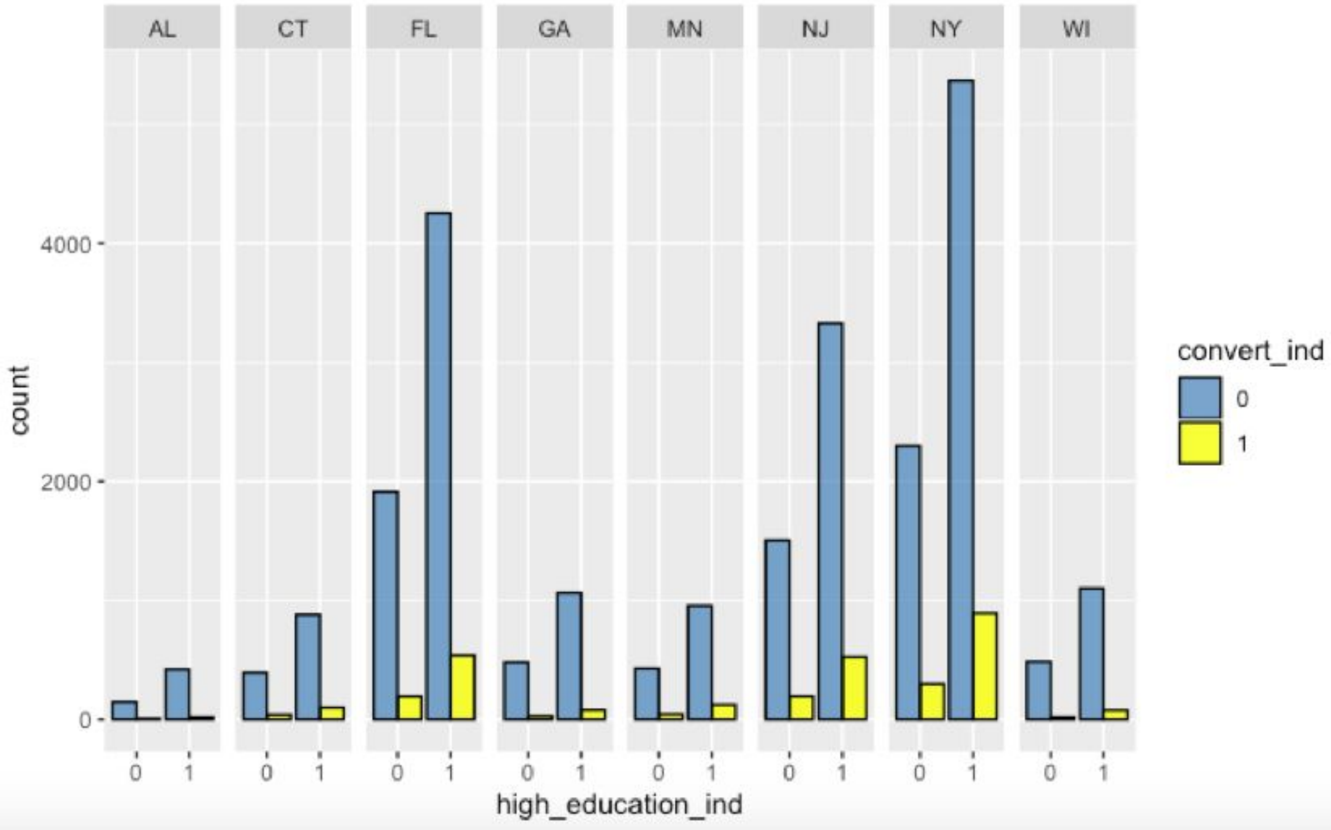
For train set for clean data:

0: 25,786

1: 3,284

Convert rate ≈ 0.127

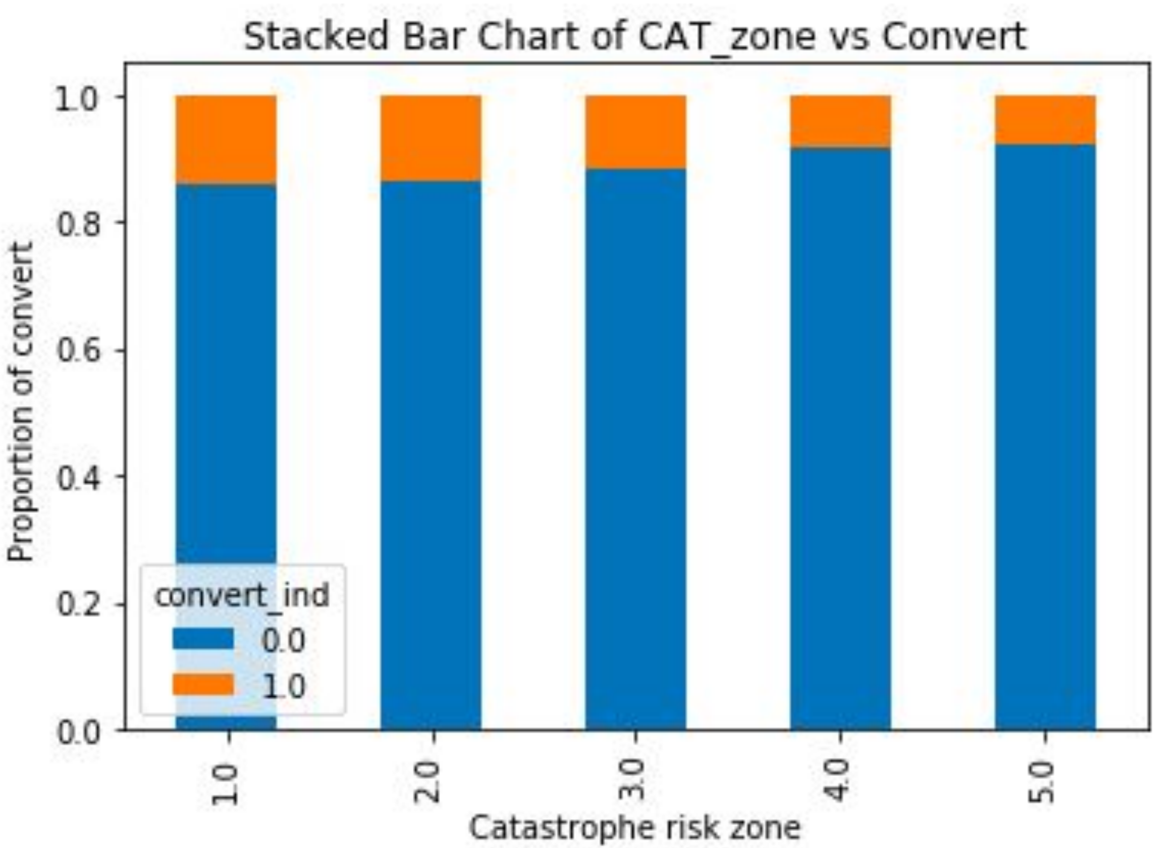
Education



Note:

Given state, more people who purchase car insurance by having higher education than lower education

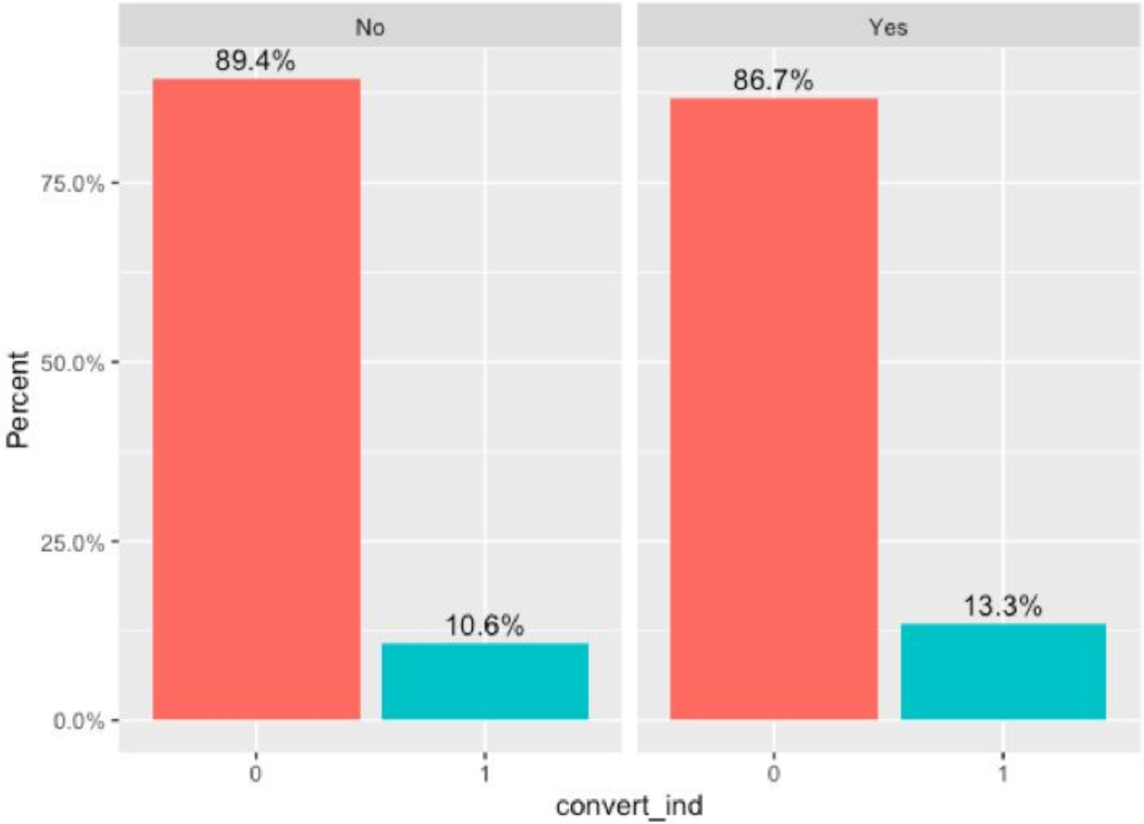
Catastrophe Risk Zone



Note:

The proportion of the customer convert depends a great deal on customer location of catastrophe risk zone

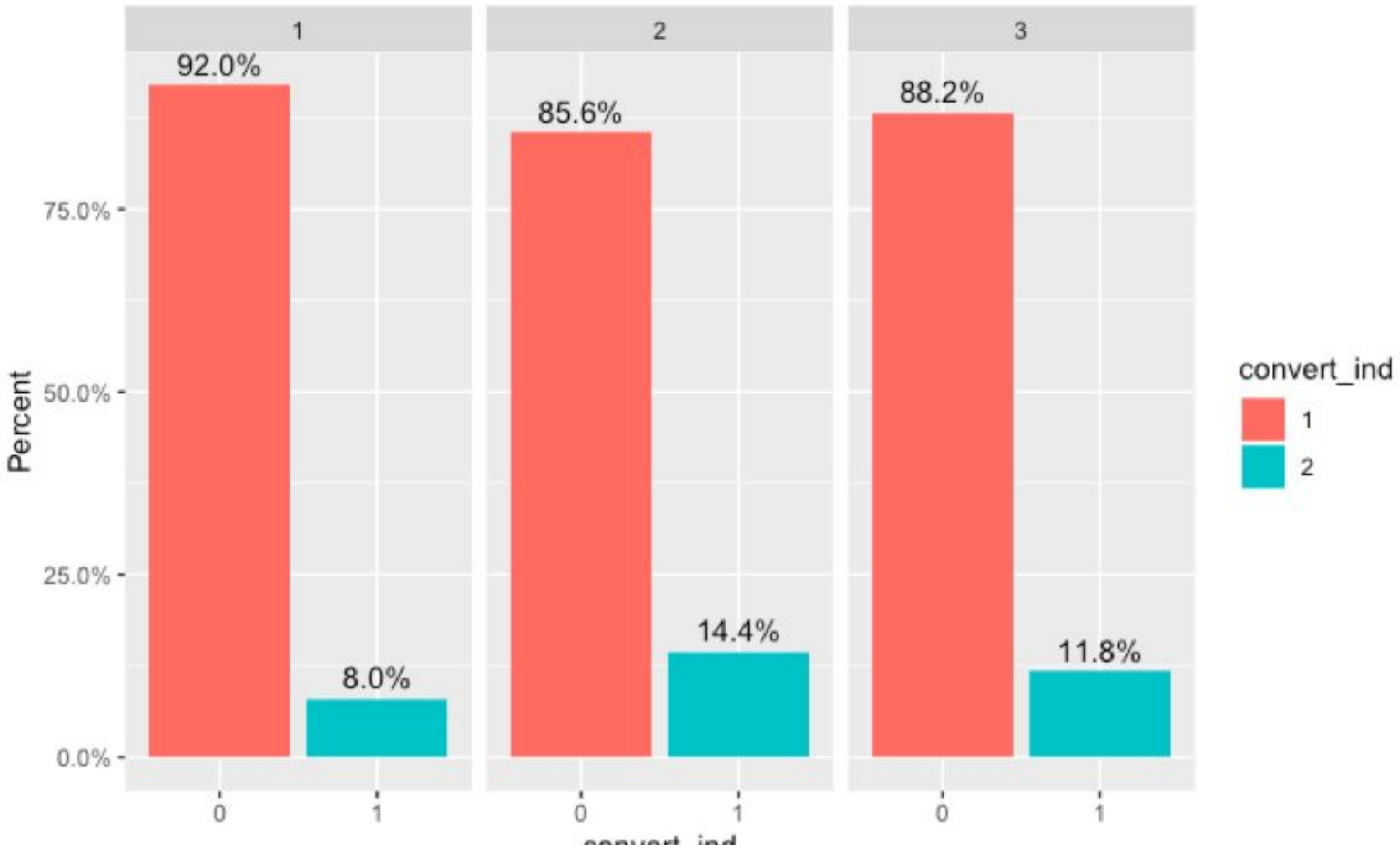
Discount



Note:

Having discount customers have higher convert percent than no discount customer

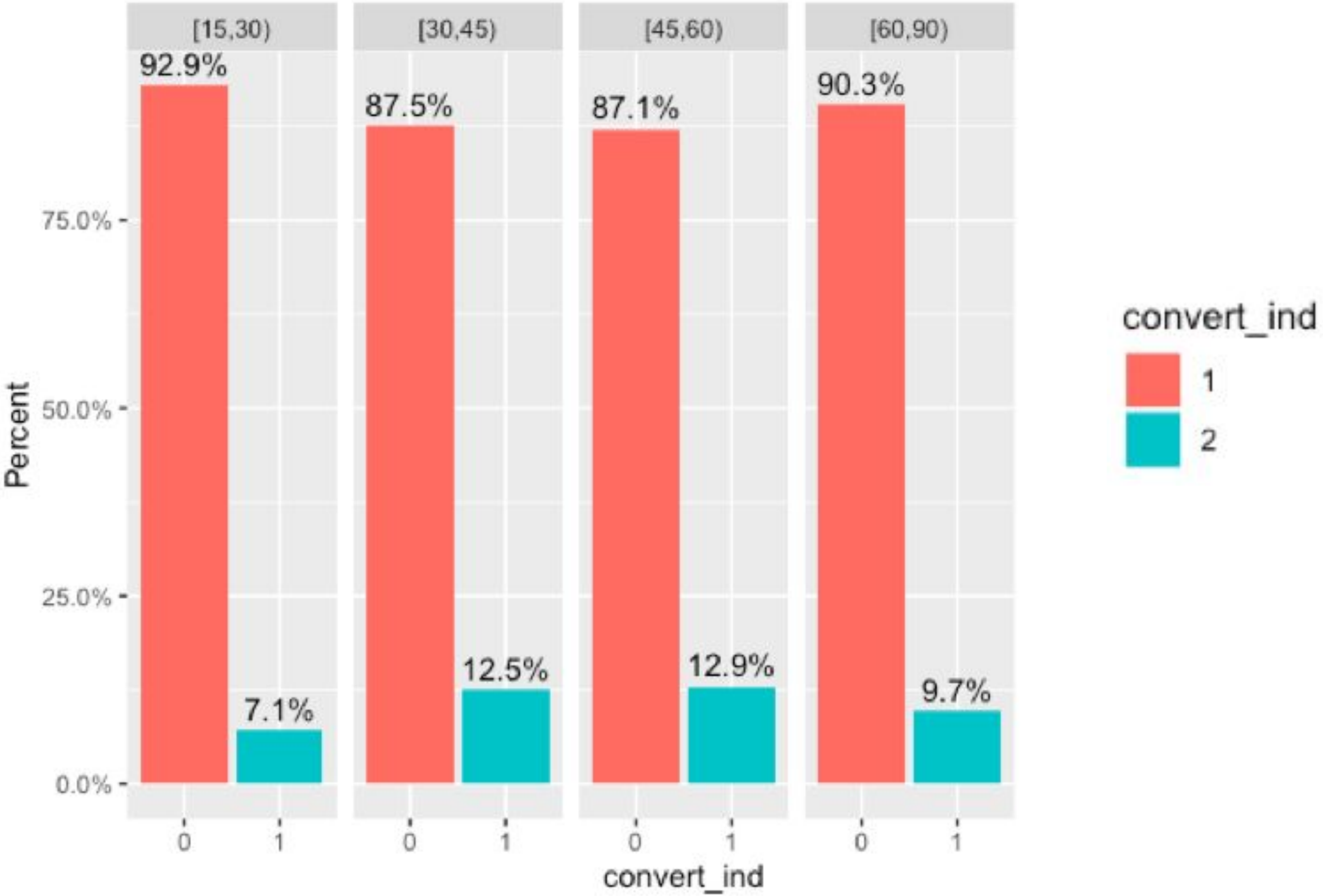
Coverage Package Type



Note:

Level 2 of packages have a little higher convert percentage

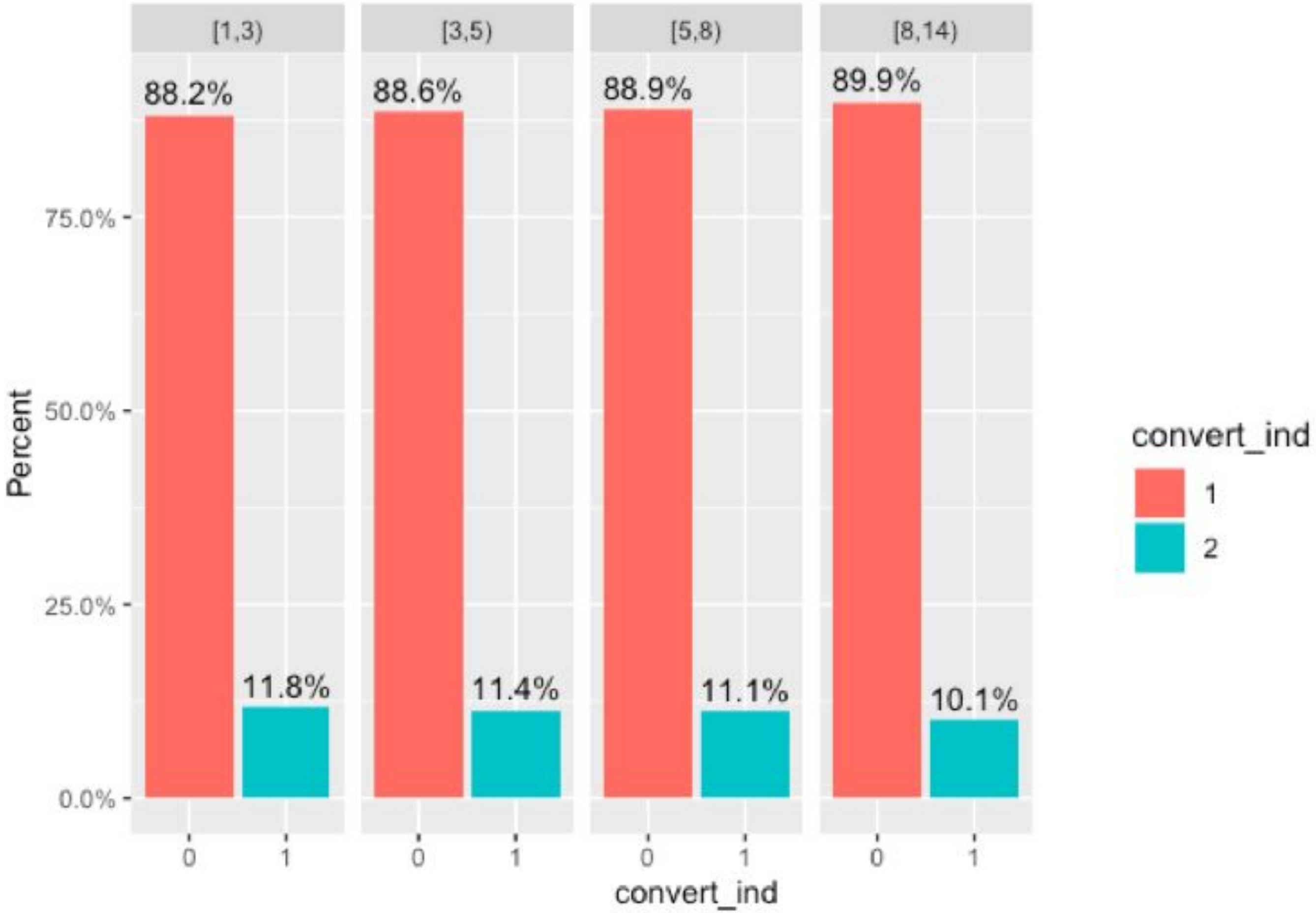
Average age of driver



Note:

Average of age between 45 and 60 has a higher percent of convert

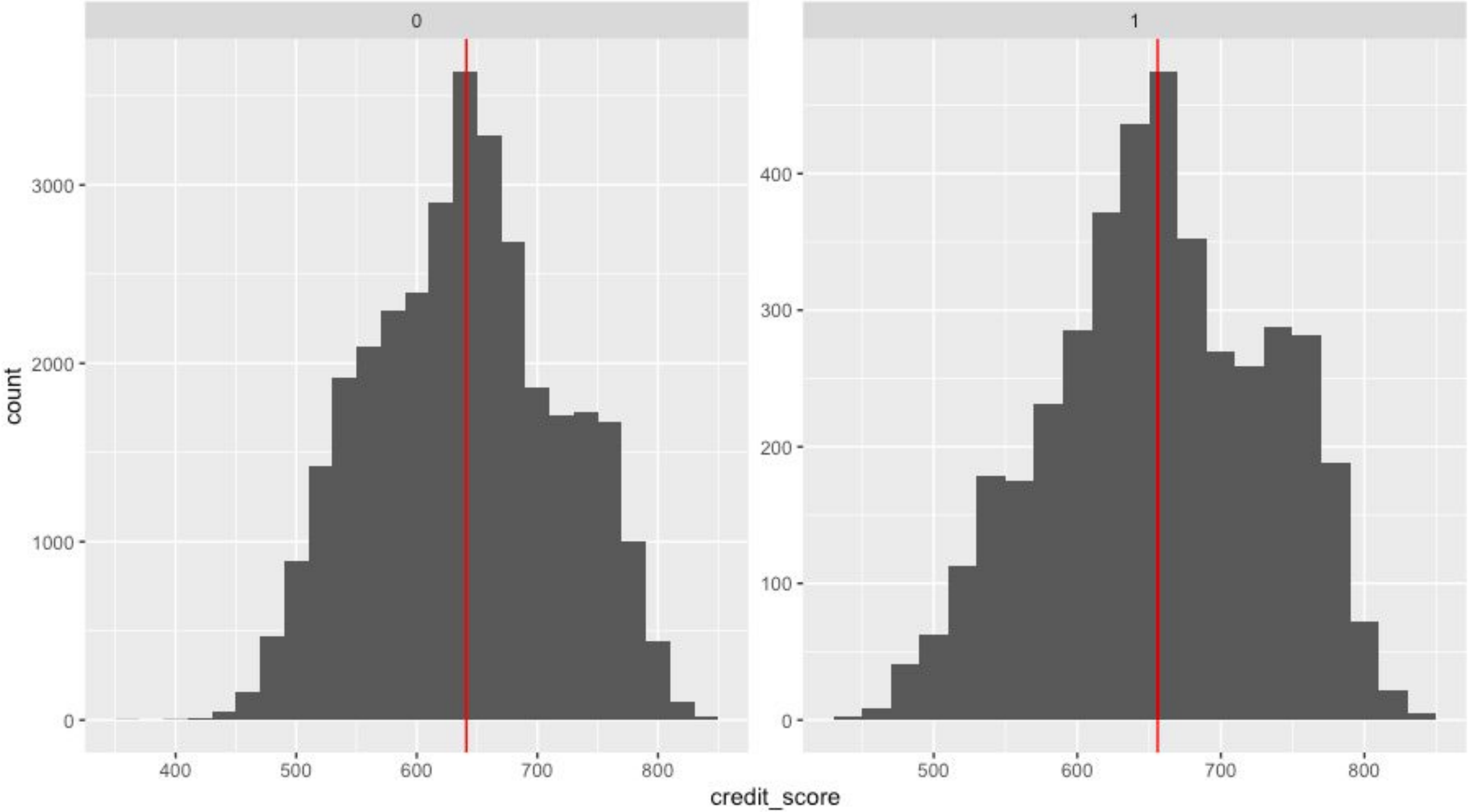
Average age of vehicles



Note:

Percent of convert decreases as average age of vehicles increase

Credit Score

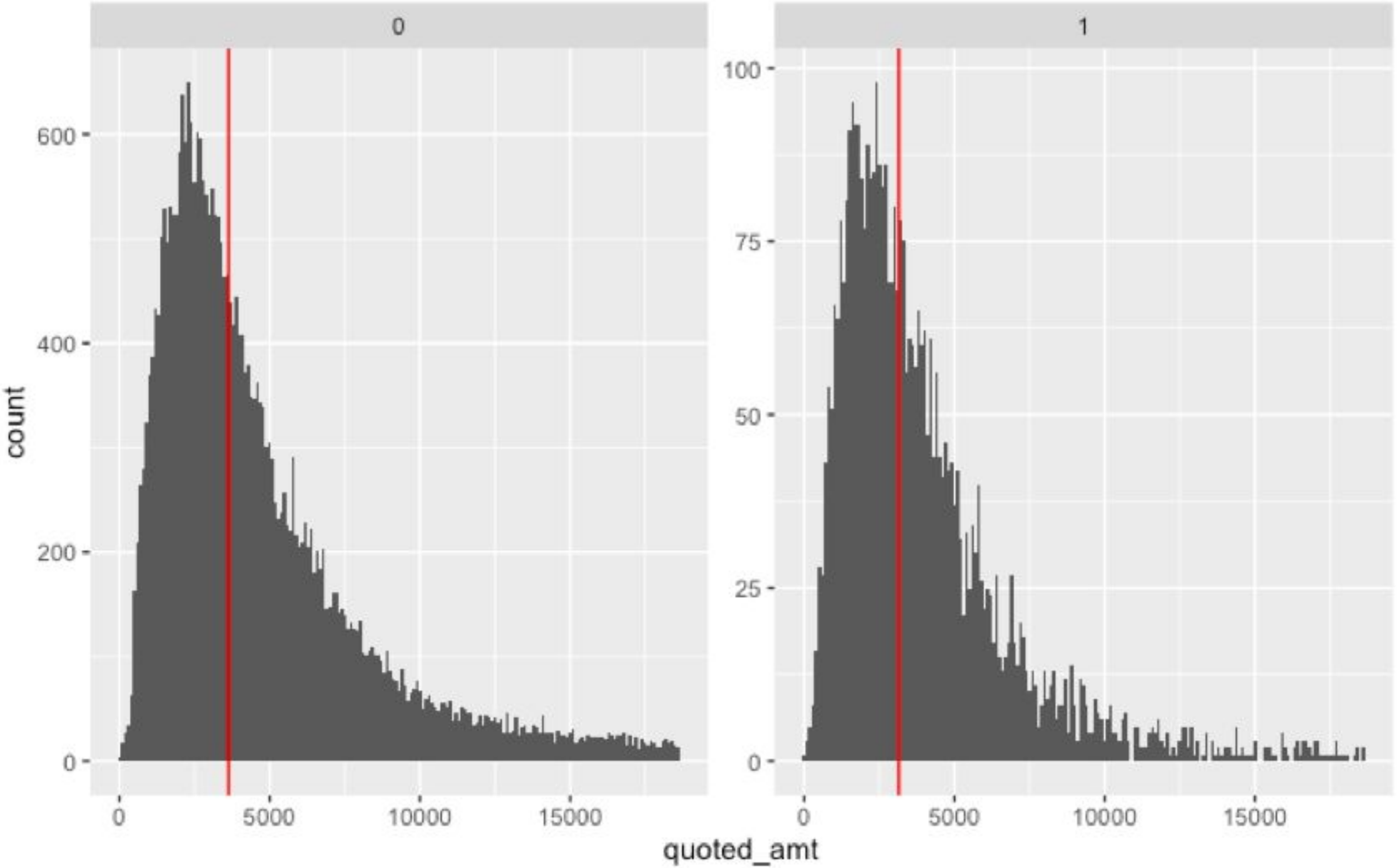


Note:

More users' s credit score are in range of 600-700

Average of credit score in converted user is little higher than unconverted users

Quoted Amount



Note:

Uncovered users quoted amount are more spread and average of amount is little higher than converted user.

Factors (not included in dataset) may help predict

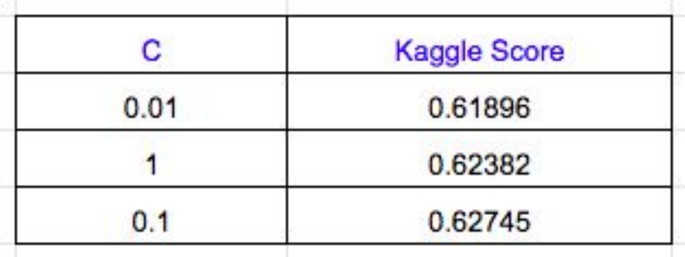
- Family income
- Population density
- Road situation
- Crime rate

Model

- ❑ Logistic Regression
- ❑ Random Forest
- ❑ KNN (K-Nearest neighbors)
- ❑ Support Vector Machine

Logistic Regression with Lasso Regularization

- Model training



C	Kaggle Score
0.01	0.61896
1	0.62382
0.1	0.62745

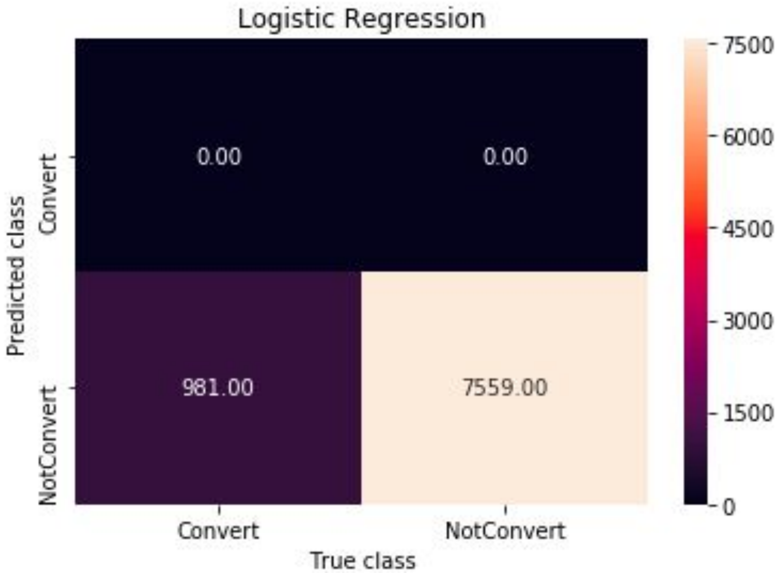
Note: C is Inverse of regularization strength.

- Result

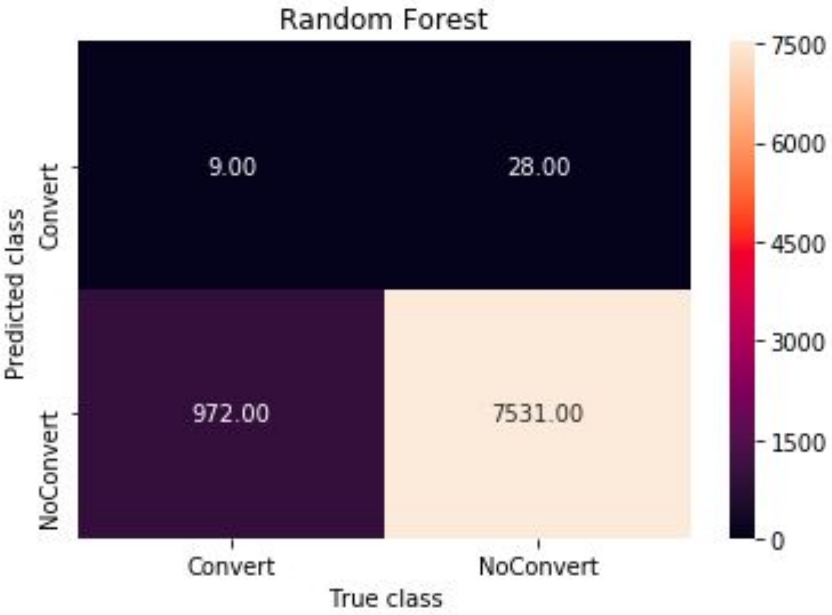
-- Accuracy: 0.885

-- Precision: 0

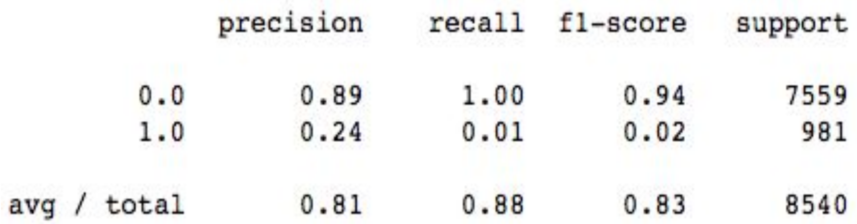
-- Recall: 0



Random Forest

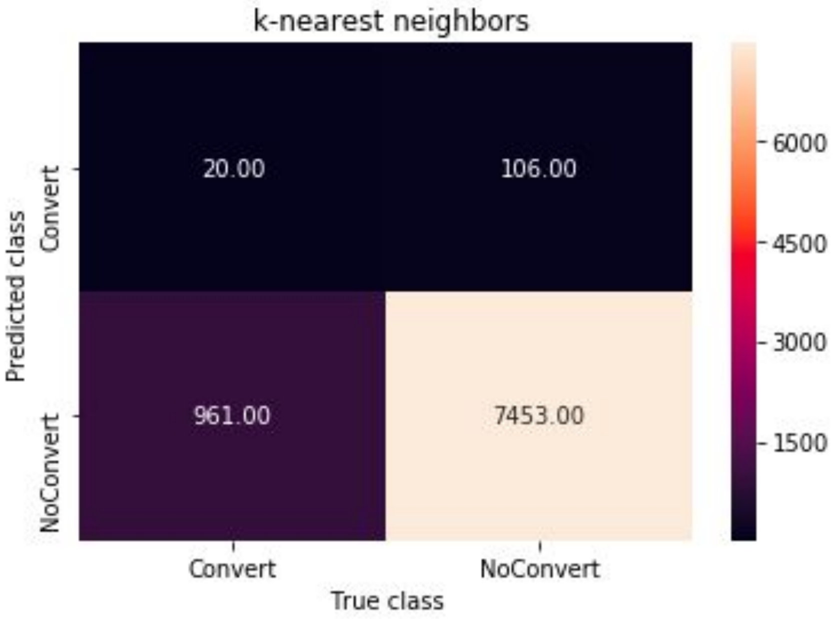


- Accuracy: 0.883
- Other performance:

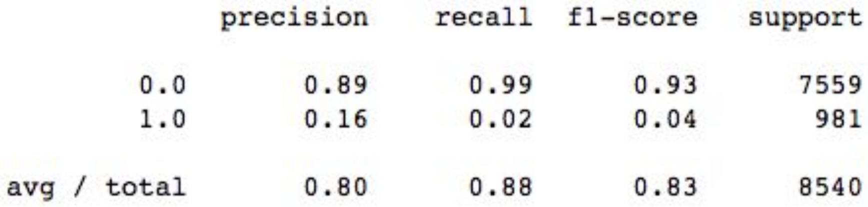


```
             precision    recall  f1-score   support

        0.0       0.89      1.00      0.94      7559
        1.0       0.24      0.01      0.02       981

avg / total       0.81      0.88      0.83      8540
```

KNN

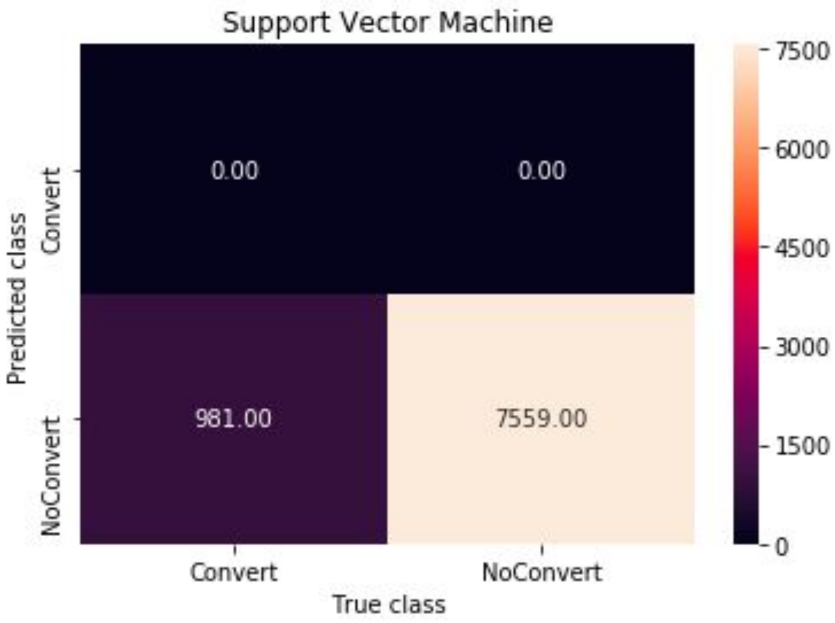


- Accuracy: 0.875
- Other performance:



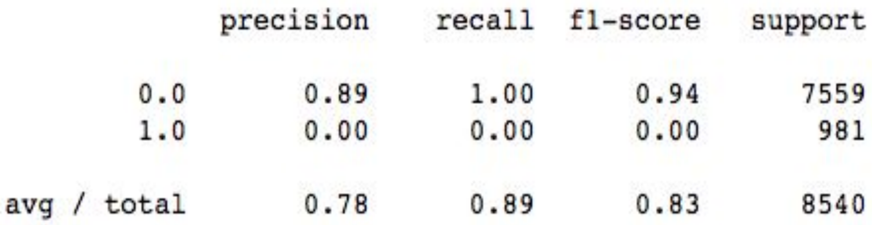
```
             precision    recall  f1-score   support

        0.0       0.89      0.99      0.93      7559
        1.0       0.16      0.02      0.04       981

avg / total       0.80      0.88      0.83      8540
```

Support Vector Machine



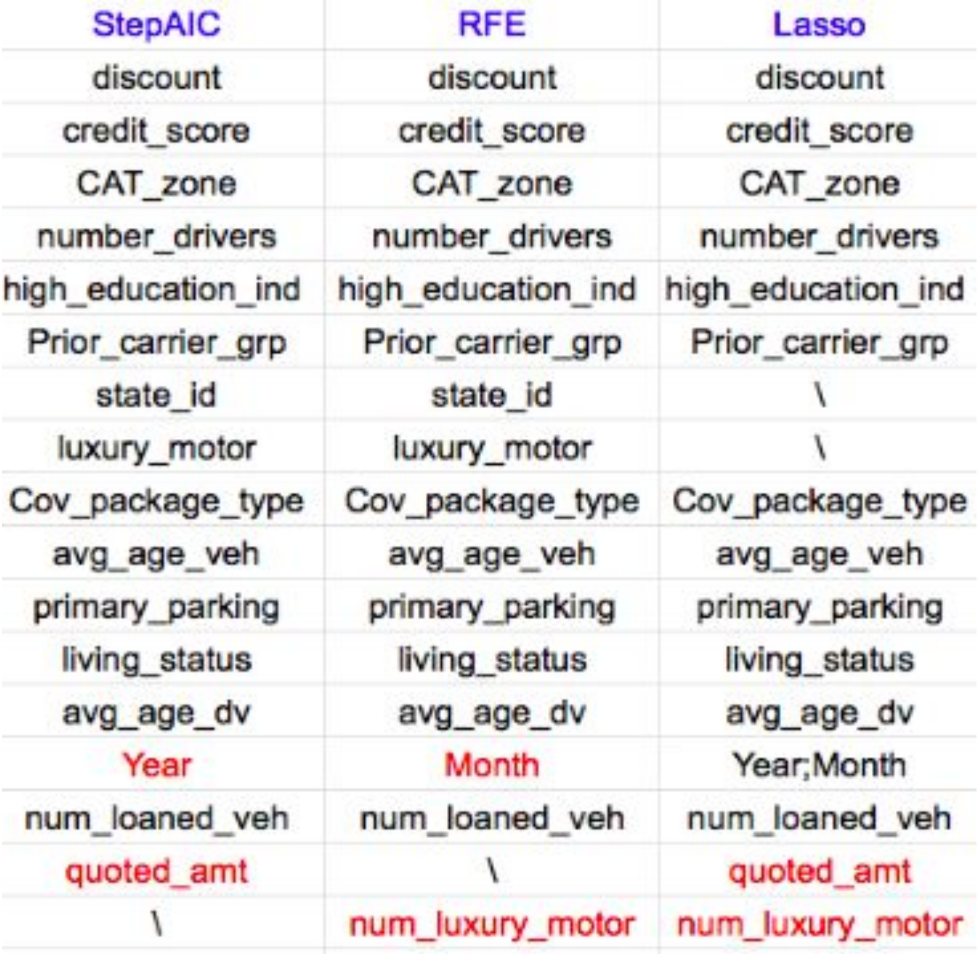
- Accuracy: 0.885

- Other performance:



```
             precision    recall  f1-score   support

        0.0       0.89      1.00      0.94      7559
        1.0       0.00      0.00      0.00       981

avg / total       0.78      0.89      0.83      8540
```

Feature Selection Result



StepAIC	RFE	Lasso
discount	discount	discount
credit_score	credit_score	credit_score
CAT_zone	CAT_zone	CAT_zone
number_drivers	number_drivers	number_drivers
high_education_ind	high_education_ind	high_education_ind
Prior_carrier_grp	Prior_carrier_grp	Prior_carrier_grp
state_id	state_id	\
luxury_motor	luxury_motor	\
Cov_package_type	Cov_package_type	Cov_package_type
avg_age_veh	avg_age_veh	avg_age_veh
primary_parking	primary_parking	primary_parking
living_status	living_status	living_status
avg_age_dv	avg_age_dv	avg_age_dv
Year	Month	Year;Month
num_loaned_veh	num_loaned_veh	num_loaned_veh
quoted_amt	\	quoted_amt
\	num_luxury_motor	num_luxury_motor

Note:

StepAIC :

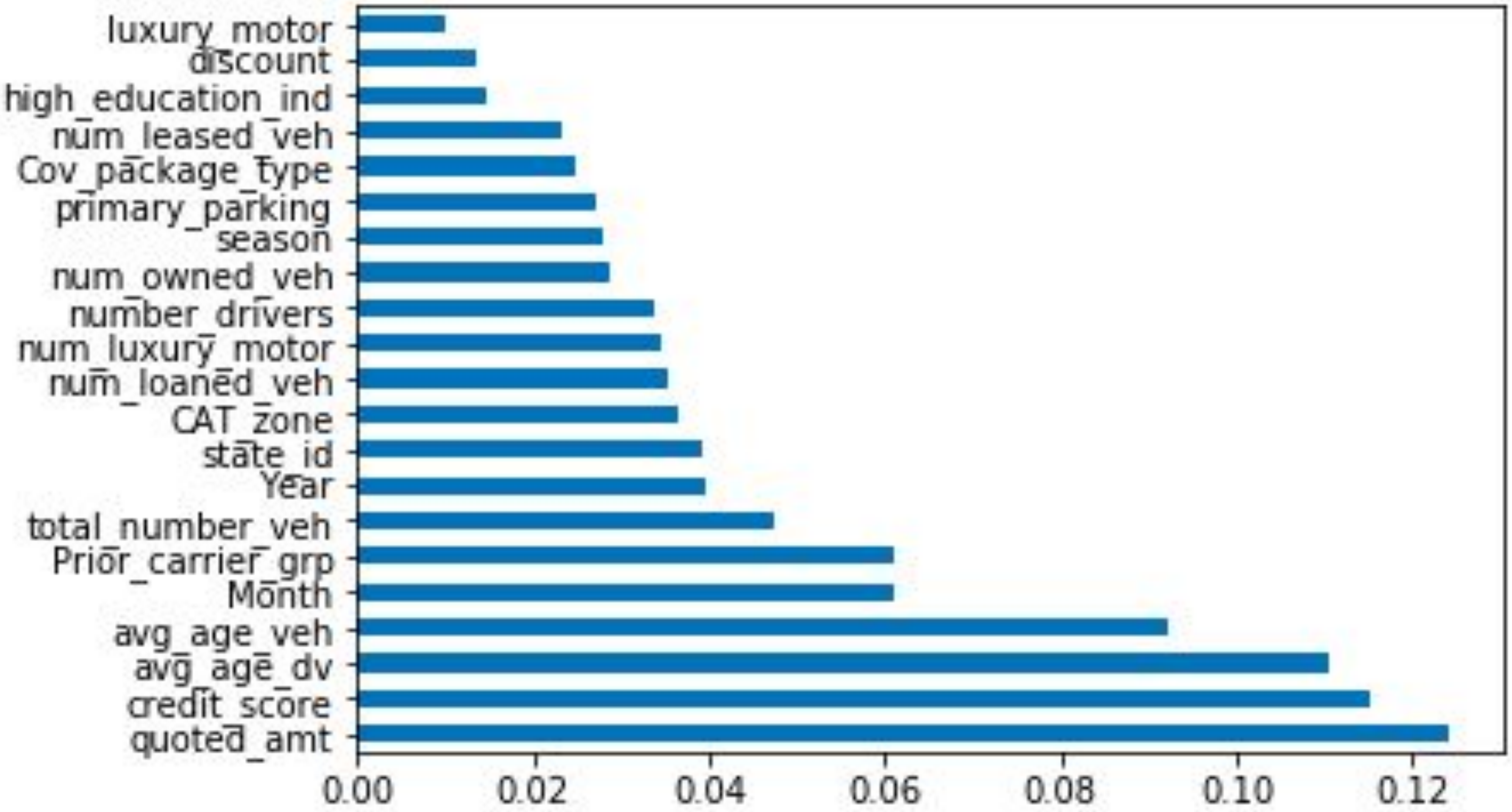
Performs stepwise selection.

(**AIC**, Akaike information criterion)

RFE :

Recursive feature elimination

Feature Importance



Recommendation

Insights and Leverage: All of selected factors are important at predicting whether someone who is convert or not.

❑ Credit_score is higher,discount is higher, education is higher,the higher probability of convert.

❑ Catastrophe risk zone, if you live in a hurricane zone or flood plain, you may need to carry insurance on your car.

❑ Cov_package_type, if Traveler make plans to target users, may look at the level 2 of package of users