STAT8051 Project Report: Travelers Statistical Modeling Competition

Jooyong Lee

lee02628@umn.edu

1 Introduction

Discerning factors that affect potential customers' decisions on paying for a service or product is significant in any business. Such factors could be internal factors, such as part of the business's strategy, or external factors, such as characteristics of potential customers. Offering insights into external factors, such as who has a higher chance to convert to customers and what characteristics of customers made this conversion, is one of the central roles of a company's data analytics team. These insights make the company concentrate on putting effort into such customers, decreasing the cost that the company should spend.

To significantly contribute to a business strategy, this project will identify quoted policies that Peace of Mind Insurance Company will convert (a.k.a. issue) and reveal factors to determine which features of customers or policies impact conversion rate. Three significant steps to proceed with the project are Exploratory Data Analysis (EDA), feature engineering, and fitting models.

2 Data Overview

As shown in Table 1, Peace of Mind Insurance Company offers three different datasets: Policies.csv, Drivers.csv, and Vehicles.csv. Every dataset includes policy_id as an attribute as a unique customer identifier. Policies.csv consists of a train and test set, and a conversion indicator is missing for policies in the test set, and a conversion model this project will suggest will predict those.

Dataset	# Samples	# Features	Features
policies.csv	49162	21	policy_id, convert_ind, Quoted_dt, discount, Home_policy_ind, zip ...
drivers.csv	106294	6	policy_id, gender, living_status, age, safty_rating, high_education_ind,
vehicles.csv	169237	6	policy_id, car_no, ownership_type, color, age, make_model

Table 1: Details of Datasets

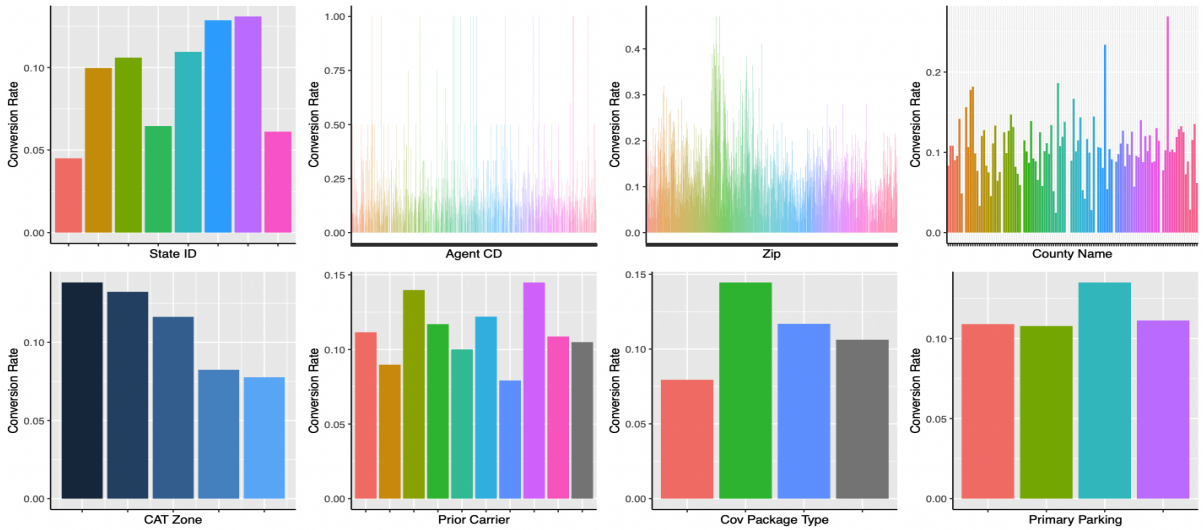


Figure 1: Exploratory Data Analysis (EDA) of Policies.csv

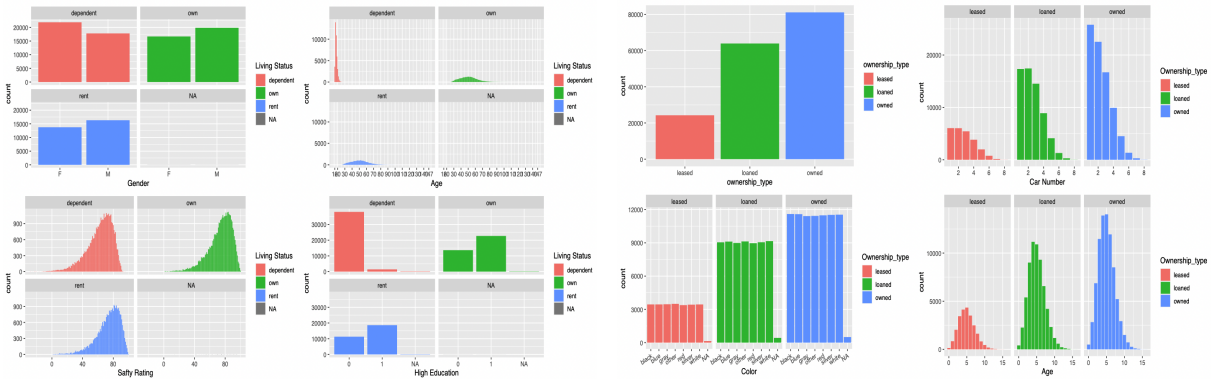


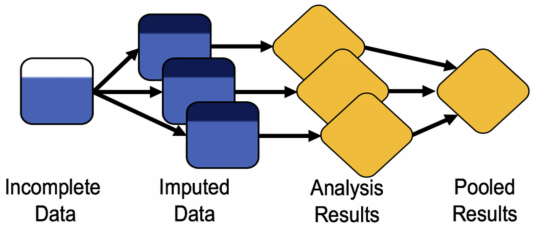
Figure 2: (a) EDA of Drivers.csv (b) EDA of Vehicles.csv

2.1 Policy Dataset

Out of all 21 features of the policy dataset, the number of features we can use to train the conversion model is 18 because policy_id is the identifier of each sample, and the split indicates the sample is assigned to the train set or test set. As shown in Figure 1, the y-value of every plot represents the conversion percentage, and the x-values vary, such as State ID, Agent CD, and Zip. Among observations in the Figure, we can observe that conversion rate and the feature CAT Zone have a negative association.

2.2 Drivier Dataset

With the Driver dataset, plots in the figure 2 count the number of samples depending on living status related to other features: Gender, Age, Safty Rating, and High Education. However, there is no unexpected observation.



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

Figure 3: (a) Steps of MICE algorithm (b) Example of converting zip

2.3 Vehicle Dataset

Similarly, as plotting the driver dataset, plots in figure 2 count the number of samples depending on the ownership type of car related to other features: the number, color, and age of the car.

3 Data Pre-Processing

3.1 Imputation

As shown in the table, some features in the policy dataset have missing values. Those missing values are imputed using Multiple Imputations with Multivariate Imputation by Chained Equations (MICE) algorithm 3. Steps to implement Mice are:

1. Separate the whole dataset into several sub datasets.
2. Fill the missing variable of each sub-dataset conditionally on all other variables.
3. The imputed datasets are each analyzed and then combined into the final result.

3.2 Converting Zip code

Due to large number of classes for feature Zip, it is converted to longitude and lattitude, and added two corresponding columns to policy dataset. Here 3 are some examples.

3.3 Merging

One dataset, including a helpful feature for predicting tasks, is necessary to train the model. The obstacle to merging datasets was that driver and vehicle datasets are 'one and many' structures. Therefore, the mean of numeric feature and mode on categorical feature is used for the corresponding features of an integrated sample. If there are many modes for the categorical feature of one policy, then the mode is calculated from the whole sample and used.

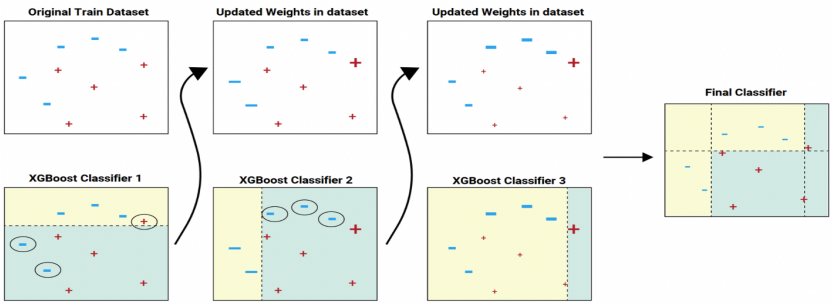


Figure 4: Process of Boosting

4 Model Fitting

Out of the whole features possibly used to train the model, 25 features, such as driver age, living status, discount, longitude, and latitude, are selected after implementing backward elimination. As shown in Table 2, three machine learning models are trained for predicting conversion status.

Models	AUC with Train	AUC with Test (30%)	AUC with Test (70%)
Logistic Regression	0.6382	-	-
Random Forest	0.6597	-	-
XGBoost	0.6838	0.67665	0.69377

Table 2: Performance of Models

4.1 Model3: eXtreme Gradient Boosting

The model shows the best performance is eXtreame Grandient Boosting (XGBoost). XGBoost is Gradient Tree Boosting with regulariztion to counter overfitting models by lowering variance while increasing some bias. The base learner of XGBoost is decision tree, and, as described in Figure 4, creates a seqeunce of models(Decision tree) that attempt to correct the mistakes of the models before them in the sequence using gradients.Following describes the algorithms to update the model to predict.

$F_o(x) = argmin_p \Sigma_{i=1}^{N} L(y_i, \rho)$
For m = 1 to M do:
$\tilde{y}_i = -[\frac{(y_i, F(x_i))}{(x_i)}]_{F(x)=F_{m-1}(x)}, i = 1, N$
$m = argmin_{\alpha,\beta} \Sigma_{i=1}^{N} [\tilde{y}_i - \beta h(x_i; \alpha)]^2$
$\rho_m = argmin_\rho \Sigma_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i; \alpha))$
$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m)$
endFor
end Algorithm

Feature	Coefficient
driver_age	2.15e-3
living_dependent	0
living_own	1.28e-1
living_rent	1.27e-1
cat zone	-2.23e-2
driver number	-2.34e-2
discount	4.48e-2
quote amount	-3.91e-6

Table 3: Feature Coefficients from Logistic Regression

5 Interpretation

As shown in Figure, the Importance of the model's feature is presented in descending order. The impressive part about this is that the Importance of longitude is about 0.052, and the Importance of latitude is about 0.031. It could be evidence that the weather affects to conversion rate. To illustrate the black box, Logistic Regression is used, and Table 3 shows the coefficient for the features that have high importance. In terms of the coefficient result, higher driver age and discount and lower cat zone, driver number, and quote amount can be suggested to make a higher probability of conversion.

6 Conclusion

This project contributed to constructing one dataset using statistical methods such as Backward Elimination, Multiple Imputation, and Boosting to train a machine learning model. As part of feature engineering, the zip code was converted to longitude and latitude, improving the model's performance. Even though the XGBoost model shows the best performance, because of a lack of interpretation, feature coefficients are calculated with Logistic Regression. As a further analysis, the feature agent_cd could be converted to another form because agent_cd has many different classes. Moreover, features used to train the model should be observed more to check if there are any correlations with each other.

7 References

https://www.sciencedirect.com/science/article/pii/S0828282X20311119
https://blog.quantinsti.com/xgboost-python/
https://www.researchgate.net/figure/The-multiple-imputation-MI-process-In-the-first-step-missing-data-shown-in-white-are_fig4_334213038