Name: Rohit Garla
Net ID-rg3365
Name: Mengrui Zhang
Net ID-mz2258

Assignment 2

Analysis of Data

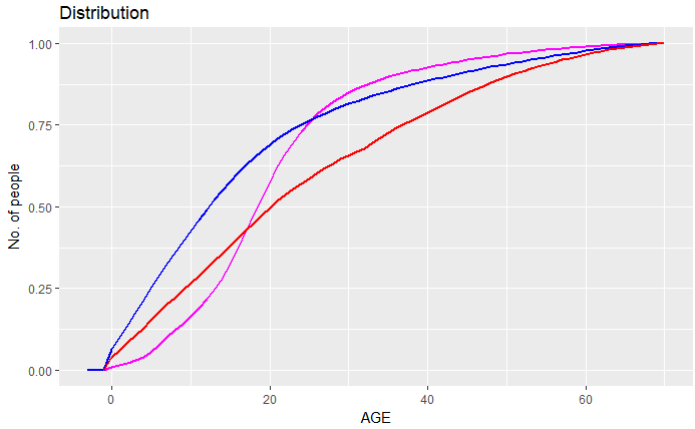
label <int>	count <dbl>
1	92996
3	27095
2	2602038

- When we grouped the data based on label, we found out that more observations belonged to the “current” category. And mark here, the imbalance of our data may cause inaccuracy of our model later.
-

FICO <dbl>	CLEANLTV <dbl>	INCENTIVE <dbl>	AGE <dbl>
716	75	0.00265	-3
683	47	0.00110	-2
764	80	0.00215	-2
656	73	0.00190	-2
744	68	0.00075	-2
773	95	-0.00780	-2
654	84	0.00255	-2
716	75	0.00345	-2
742	35	0.00130	-2
764	95	0.00465	-2

1–10 of 1,000 rows

- On filtering the data based on AGE
-



- From the graph we can observe that cumulative distribution of No. of people are different for different group. This is also an intrinsic motivation that our AGE parameter may be valid, since it does have difference patterns for different groups. We can do this analysis for other variables later.

Analysis of Model

Logistic Regression Model

For the assignment, we have used multinomial logistic regression, since we are dealing with dependent variable which has three categories.
The underline technique will be same as the logistic regression for binary classification until calculating the probabilities for each target. Once the probabilities were calculated. We need to transfer them into one hot encoding and uses the cross-entropy methods in the training process for calculating the properly optimized weights.

Inputs

The inputs to the multinomial logistic regression are the features we have in the dataset. Since, we are going to predict the label, the features will be incentive, age, FICO etc. These features will treat as the inputs for the multinomial logistic regression.

Model Co-efficient and Intercept

```
Multinomial coefficients: DenseMatrix([[ -8.96215336e-04,  -1.56239975e-03,  -8.87493146e-01,
               -3.75400567e-04,  -5.20023376e-03,  -4.80593573e-03,
               -4.78263121e-03,  -2.61713366e-03],
              [  1.72915021e-03,  -1.32352183e-02,   3.76449456e+01,
               -9.17912597e-03,  -6.06413175e-02,  -3.99191287e-01,
               -1.64370725e-01,   1.59645654e-02],
              [  1.89525143e-03,  -7.31940466e-03,  -6.89399735e+01,
                1.88828396e-05,   2.43537855e-02,   2.76796727e-01,
                6.94222871e-02,   3.86378857e-03],
              [ -2.72818630e-03,   2.21170227e-02,   3.21825211e+01,
                9.53564370e-03,   4.14877658e-02,   1.27200496e-01,
                9.97310687e-02,  -1.72112203e-02]])
Multinomial intercepts: [-5.29636850861,2.13578062824,3.78207362107,-0.621485740706]
```

- The model built is a multinomial regression model, so it will have co efficient as co efficient matrix and intercepts as an intercept vector.

Evaluation Matrix

- On Evaluating the logistic regression model with MulticlassClassificationEvaluator we achieved a score of 0.7189
- The Entropy for the Logistic Regression model was 0.2242

Training data input volume for different groups.

label	count
1	1864
2	8033
3	103

Random Forest Model

A random forest is a meta estimator that fits many decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. Random forests are ensembles of decision trees. Random forests are one of the most successful machine learning models for classification and regression. They combine many decision trees to reduce the risk of overfitting. Like decision trees, random forests handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and are able to capture non-linearities and feature interactions. Random forests train a set of decision trees separately, so the training can be done in parallel. The algorithm injects randomness into the training process so that each decision tree is a bit different. Combining the predictions from each tree reduces the variance of the predictions, improving the performance on test data. The test error is a measure for calculating the accuracy of Random forest.

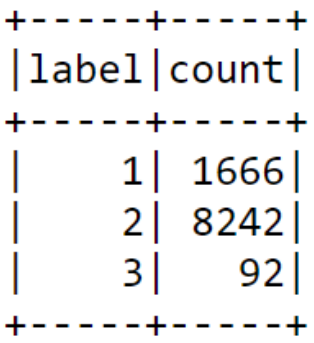
Higher than true value(H>1.9)
- The entropy obtained here was 0.2042
- The test error for this model was 0.2357

- Training data input volume for different groups:

label	count
1	1532
2	8386
3	82

Lower than true value(H<1.9)
- The entropy obtained here was 0.1884
- The test error for this model was 0.2204

- Training data input volume for different groups:



label	count
1	1666
2	8242
3	92

Conclusion

Based on our models, we may find that the input volume may extremely affect our result of entropy. It seems that, this should have some skewness of our real group volume. So that if we don't stratified subsample our dataset, we will get a very bad result of entropy. For example, if we just take the 10000 observations from dataset and use them to train our model, the entropy will be 0.05, which is extremely far from our real value 0.19. As a result, we can conclude it is a under fit. However, if we try to equally assign volume for each group, we may cause an overfitting, and the entropy h may become 0.9. Therefore, it is crucial to decide what's the true ratio of there different groups.

Question:

1. How to do a features selection? Our group want to try to change the model by changing the feature items, however, I don't know how to do that.
2. How to change a double array to a sql vector. Since I find that model only allow to use sql vector as features.