Analytics Edge Project Report

Students: Wang Yiran (1000906), Zhou Jiuqi (1000904)

Data Feature pre-processing

Before we started fitting the models, we firstly pre-processed all the train and test data sets. This was for more convenience and accuracy when constructing models.

Factorial variables to numeric variables. The first thing we did is to change factorial variables into numeric variables by their meaning. Here are the details of how we did this:

Factorial variables	Numeric variable
gender: Male, Female	*gend*: 1, 0
segment: Full-size Pickup, Midsize Car, Midsize Utility, Midsize Luxury Utility segments, Prestige Luxury Sedan, others	*car*: 1, 2, 3, 4, 5, 6
ppark: Never, Daily, Monthly, Weekly, others	*newpark*: 0, 365, 12, 52, 1
income: Under $29,999, $30,000 to $39,999, $40,000 to $49,999, $50,000 to $59,999, $60,000 to $69,999, $70,000 to $79,999, $80,000 to $89,999, $90,000 to $99,999, $100,000 to $109,999, $110,000 to $119,999, $120,000 to $129,999, $130,000 to $139,999, $140,000 to $149,999, $150,000 to $159,999, $160,000 to $169,999, $170,000 to $179,999, $180,000 to $189,999, $190,000 to $199,999, $200,000 to $209,999, $210,000 to $219,999, $220,000 to $229,999, $230,000 to $239,999, $240,000 to $249,999, $250,000 to $259,999, $270,000 to $279,999, $280,000 to $289,999, $290,000 to $299,999, others	*money*: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30
educ: High School, Some College (1-3 Years), College Graduate (4 Years), Grade School, Postgraduate College, others	*educa*: 1, 3, 4, 5, 6, 2
region: MW, NE, SE, SW, others	*place*: 1, 2, 3, 4, 5
Urb: Rural/Country, Suburban, others	*liveplace*: 1, 2, 3
miles: Under 50 Miles, 51 To 100 Miles, 101 To 150 Miles, 151 To 200 Miles, 201 To 250 Miles, 251 To 300 Miles, 301 To 350 Miles, 351 To 400 Miles, others	*distance*: 1, 2, 3, 4, 5, 6, 7, 8, 9
night: Under 10%, 10% To 20%, 21% To 30%, 31% To 40%, 41% To 50%, 51% To 60%, 61% To 70%, 71% To 80%, 81% To 90%, others	*sleep*: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
age: 30 To 39, 40 To 49, 50 To 59, 60 & Over, others	*old*: 3,4, 5, 6, 2

Create new variables by multiplying price variable. For all the cases (people) and their 19 tasks, the corresponding variables like ***car***, ***newpark, money*** are all the same. This would always cause us singular problem when creating the models. So we changed them by multiplying ***price*** variables.

Create new variables. By examining the real meaning of each variable, we constructed two new variables in the models. Firstly, we updated variable ***money*** = ***price*** / ***income***. This variable gave us the ratio of how much percentage of the price account for the income of that person. We thought it was useful to our models because it showed people's purchasing power which is critical when they made choices. Secondly, we updated a variable night = NV * sleep. NV is the variable about night vision system which we strongly believed that it's related to the time of night driving.

Modelling Methods

We have tried out three methods: multinomial logit regression, logit regression, mixed logit model, and random forest. In mlogit data set, there are total 31 features (*CC, GN, NS, BU, FA, LD, BZ, FC, FP, RP, PP, KA, SC, TS, NV, MA, LB, AF, HU, Price, age, gender, educ, distance, night, region, Urb, ppark, money, year, car*) with varied range of factor values. For example, CC takes value from 1, 2, 3, 4; while year values from 2000 to 2006. We have normalised all feature values with zero mean and unit variance as all values are in the same range from 0 to 1.

Cross Validation. Since the same dataset ("train.csv") supplied training and testing samples, 10-fold cross validation was used and 90% of the dataset was selected for training in each fold. The cross validation is utilized to prevent the circumstance that

the model was overfitting on the train set with large bias on the test set. The valuation criterion was the average of log likelihood from 10 times of prediction the same as the online criteria.

Multinomial Logit Regression Model. At the beginning, we have tried out the multi-logit regression model by adding different additional features created in the first part into the original attributes set. The original attributes set is *CC, GN, NS, BU, FA, LD, BZ, FC, FP, RP, PP, KA, SC, TS, NV, MA, LB, AF and HU.* The following table (*Table 1*) shows the cross validated log likelihood of each model. Model 1 was the basic model with all original features while model 2 has included all new features. In the summary of model 2, it indicated that *CC, price, night, year*, and *car* were not significant at 1 level and *distance* was only significant at 0.1 level. Hence, in the following models, we have tried out to get rid of those insignificant variables except price one each time. Because the we suspected that the *price* was insignificant due to the overfitting of the new added features that was operated with *price*. The best feature set for the multi-logit regression model is the combination in model 12.

Table 1. mLogit Regression Cross-Validation Log Loss with Additional Features that Multiplied with *Price*

Additional Features	age	gender	educ	distance	night	region	Urb	ppark	money	year	car	Original: CC	Multi-Logit Lo Loss
Model 1	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✔	1.197947
Model 2	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	1.187548
Model 3	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✖	1.187534
Model 4	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✖	✖	1.187136
Model 5	✔	✔	✔	✔	✔	✔	✔	✔	✔	✖	✖	✖	1.186864
Model 6	✔	✔	✔	✔	✔	✔	✔	✔	✔	✖	✔	✖	1.187274
Model 7	✔	✔	✔	✔	✖	✔	✔	✔	✔	✖	✔	✖	1.186756
Model 8	✔	✔	✔	✔	✖	✔	✔	✔	✔	✖	✖	✖	1.186359
Model 9	✔	✔	✔	✖	✖	✔	✔	✔	✔	✖	✔	✖	1.186035
Model 10	✔	✔	✔	✖	✔	✔	✔	✔	✔	✖	✔	✖	1.186511
Model 11	✔	✔	✔	✖	✔	✔	✔	✔	✔	✖	✖	✖	1.186105
Model 12	✔	✔	✔	✖	✖	✔	✔	✔	✔	✖	✖	✖	1.185643

Table 2 showed the cross-validation log loss results of the multi-logit regression models with new manipulated features: *night* and *money*. (night = NV*sleep; money

= price/income). The new manipulated feature, *money* had improved the log loss value by a bit. And the new best result was model 17 with a different features combination.

Table 2 mLogit Regression Cross-Validation Log Loss with New Manipulated Features *(money & night)*

Additional Features	age	gender	educ	distance	night	region	Urb	ppark	money	year	car	Original: CC	Multi-Logit Lo Loss
Model 13	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	1.186792
Model 14	✔	✔	✔	✔	✔	✔	✔	✔	✔	✖	✖	✔	1.186185
Model 15	✔	✔	✔	✔	✔	✔	✔	✔	✔	✖	✖	✖	1.186177
Model 16	✔	✔	✔	✔	✖	✔	✔	✔	✔	✖	✖	✖	1.186198
Model 17	✔	✔	✔	✖	✔	✔	✔	✔	✔	✖	✖	✖	1.185503
Model 18	✔	✔	✔	✖	✖	✔	✔	✔	✔	✖	✖	✖	1.185549
Model 19	✔	✔	✔	✖	✖	✔	✔	✔	✔	✖	✔	✖	1.185963

Mixed Model with Logit Regression. We were thinking of using the logit regression model on clustering cases based on their own personal features: *age, gender, educ, distance, night, region, Urb, ppark, money, year, car.* We have used the pre-processed *income* value for *money* due to the additional factor levels in test data set, and all the rest features are factors except *year* is numeric. We have run logit regression model on the probability for one case ID choosing one certain choice. The probability results on four choices were the same for one case ID across all 19 tasks, which were used as the prior probability for the later prediction. The final prediction probabilities on four choices are the linear combination of logit regression clustering (P_l) and multi-logit regression probability (P_{ml}):

$P = \alpha P_{ml} + (1-\alpha)P_l$, where $\alpha = 0.9$ performed best prediction in our experiments. Table 3 showed the log loss value improvement in the mixed model prediction. Model 17 still had the best score, and its prediction accuracy was 49.4% on the training set.

Table 3 Comparison of Cross-Validation Log Loss between Multi-Logit Regression and Mixed Models

	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 15	Model 16	Model 17	Model 18	Model 19
Multi-Logit Log Loss	1.187274	1.186756	1.186359	1.186035	1.186511	1.186105	1.186177	1.186198	1.185503	1.185549	1.185963
Mixed Model Log Loss	1.181189	1.180802	1.180511	1.180324	1.180667	1.180368	1.180498	1.180499	1.180083	1.180108	1.180419