EE517 Project: Whether people will pick up the nickel when they saw it

Wenhai Zhu wenhaizh@usc.edu
Chenxi Zhang chenxi.zhang@usc.edu

Summary:

We found that it is possible to estimate the probability that an individual will pick up the nickel on the ground when he/she saw it by using logistic regression. The place we put the nickel is the first stair in front of Doheny library. We have collected data from 2pm to 5pm during April 8th to April 21th, and from 10pm to 6pm during April 24th to April 27th. We mainly record an individual who would go through the stair, no matter he is entering or leaving the library. There are twelve variables, which are gender, age, race, in suit, with backpack, holding something, doing something, running or not, with friends, strangers around, direction, and which side of the stair. Using with friends, doing something, and direction, we built a significant model.

Problem Description:

If there is a nickel in front of you, will you see it and pick it up? What may affect the individual to pick it up or not? It depends on many factors, like age, gender, dressing, and so on. Those questions trouble us too much. So we decide to collect data to analyze which variables influence the result.

Data Collection:

The goal of this project was to determine which variables influence the probability that a given individual would pick up the nickel. As Doheny library is one of the most important libraries in the USC, it is easy to collect data we need. We set a camcorder on the left side of the second floor of the Doheny(Figure 1), and took videos from 2 pm to 5 pm, Sunday to Saturday.

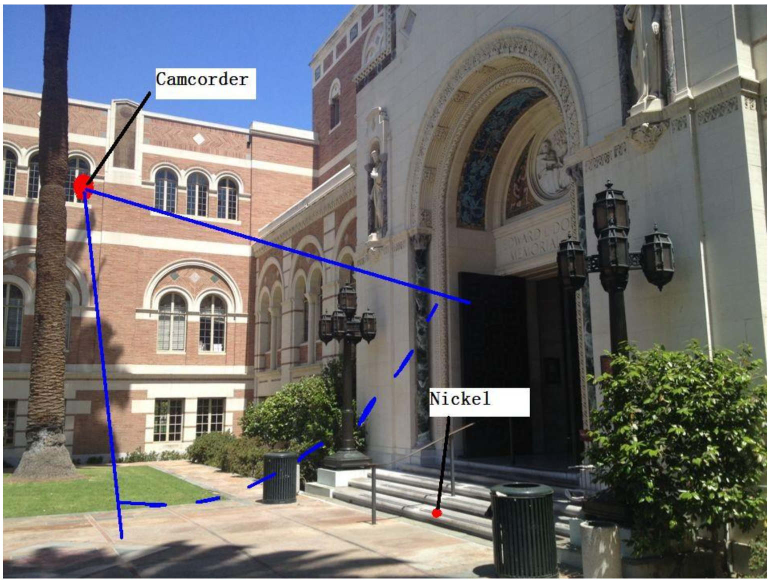

Figure 1 The location of camcorder

We divided the stair into two parts - left and right. We put the nickel on the first step of the right side when face to the main entrance. We didn't change location for both of the camcorder and the nickel to decrease the influences caused by the locations. We only considered the area in front of the Doheny Memorial Library. If an individual is neither enter the library, nor exit the library, we didn't count him as an effective data.

We set up 12 independent variables. The name and the values of the variables are as below. Figure 2 is an example.

- Gender. Contain male and female.
- Age. Contain young, middle, and old.
- Race. Contain white, black, and yellow.
- Suit. Contain in suit and not in suit.
- Backpack. Contain no backpack, one-strip backpack, and two-strip backpack.
- Empty hands. Contain no hands are empty, one hand is empty, and two hands are empty.
- Doing something. Contain doing something and not doing something.
- Running. Contain running and not running.
- Friends. Contain along, and with friends.
- Stranger. Contain no strangers and with strangers.
- Direction. Contain in and out.
- Side. Contain left side and right side.

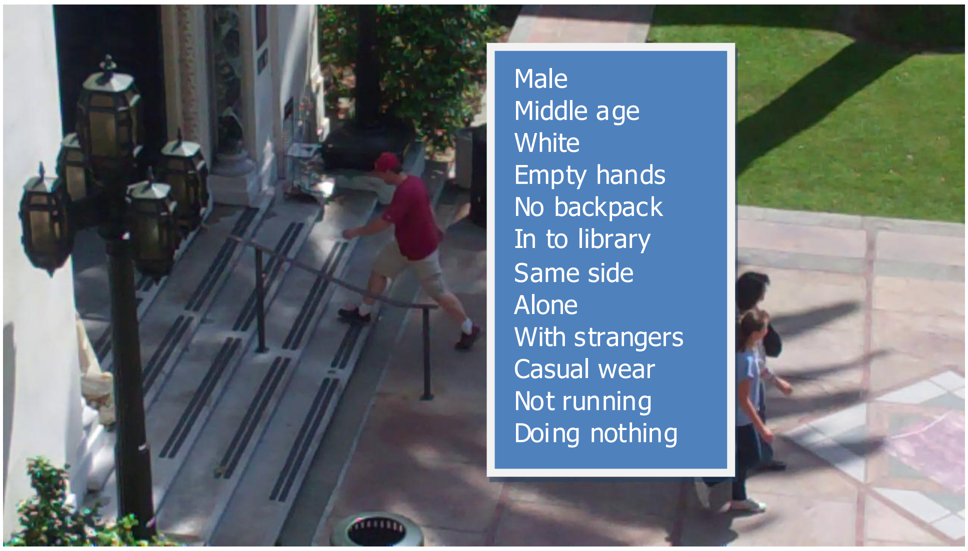

Figure 2 Example

As we need to estimate the probability of picking up a nickel in the condition of this individual must notice it first, we followed five rules (Figure 3) when we collecting data.

Rule 1: We don't count the individual who is out of "middle area"(Figure 4).

Rule 2: We don't count the individual who turns his/her head to other direction.

Rule 3: We don't count the individual who is doing his/her business and don't even notice the nickel is there. Like he/she is speaking on the phone, or reading a book.

Rule 4: We don't count the individual whose sight is blocked by other people.

Rule 5: We count the individual that his/her sight's direction is the nickel.

Rule 6: We count the individual who has obvious action that indicates he/she saw the nickel.

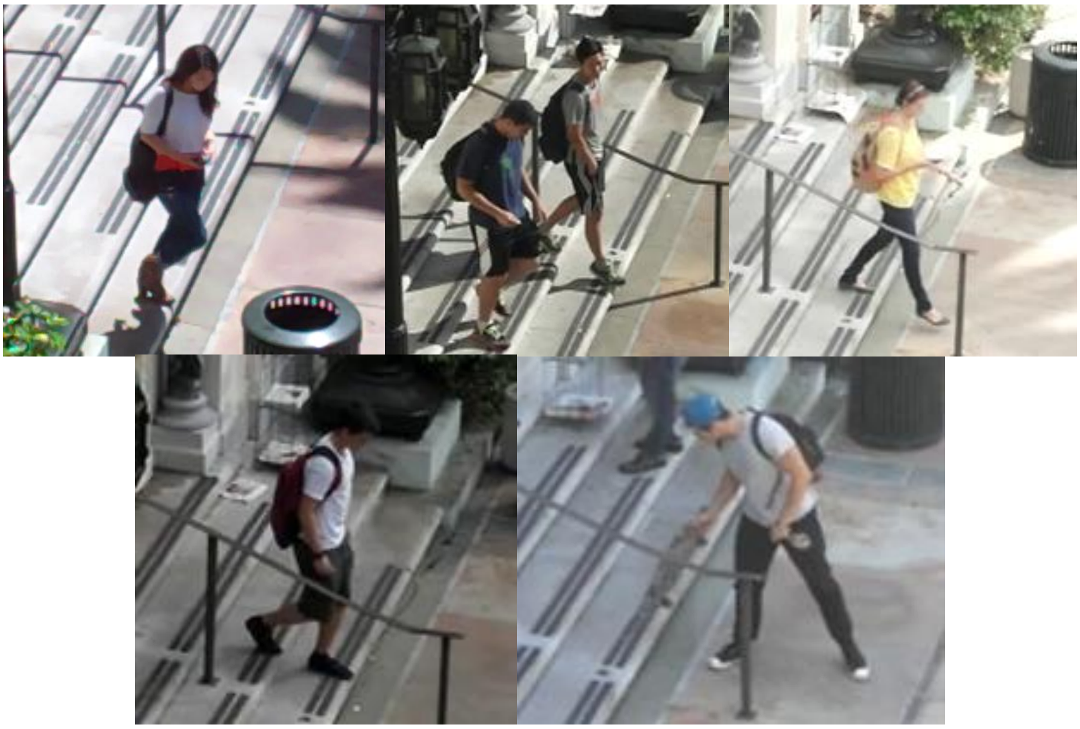

Figure 3 Examples of the five rule

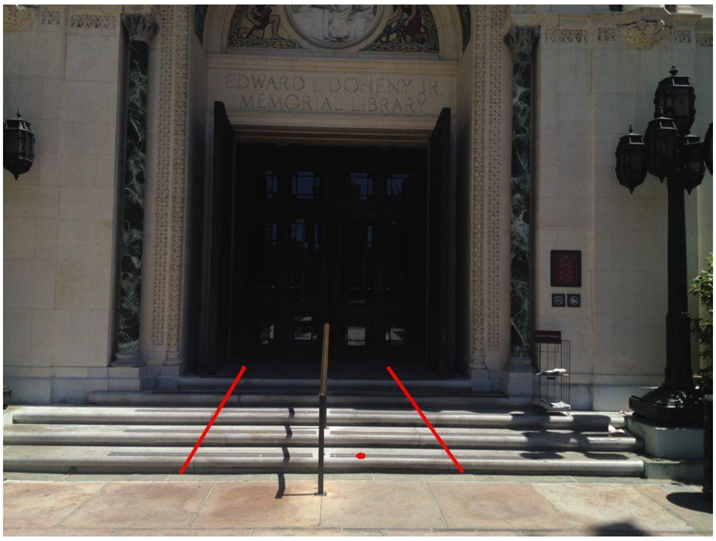

Figure 4 Definition of “Middle Area”

Logistic Regression

Logistic Regression: is a type of probabilistic statistical classification model. It is usually used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. Logistic function is as below, and t is the linear function of various variables.

$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Forward and backward stepwise: When the forward method is employed the computer begins with a model that includes only a constant and then adds single predictors to the model based on a specific criterion. While backward stepwise begins the model with all predictors included. The computer then tests whether any of these predictors can be removed from the model without having a substantial effect on how well the model fits the observed data.

Omnibus Test is a likelihood-ratio chi-square test of the current model versus the null model. The significance value of less than 0.05 indicates that the current model outperforms the null model.

Log likelihood use the observed and predicted values to assess the fit of the model. The result is negative, so we usually multiply it by -2. Then it has chi-square distribution.

Cox & Snell R square is based on the log-likelihood of the model and the log-likelihood of the original model. But it never reaches 1. So the Nagelkerke R square refined Cox & Snell R square and could reach 1.

Hosmer and Lemeshow Test is to analyze the null hypothesis that there is no difference between the observed and model-predicted value. It is used to assess the overall model.

Wald statistic is used to decide whether the coefficient is significantly different from 0.

Multicollinearity may exist in many models. The ways to detect it are: if VIF (Variance inflation factor) is more than 10, if standard error is big, if condition index bigger than 10, if eigenvalue is almost 0, then there must be multicollinearity. We also need to look at correlation matrix, if there are some numbers close to 1, then there is the multicollinerity.

Residual: Cook’s distance is the effect of deleting a given observation. It should be less than 1. The leverage is a way t detect the observations that are far away from corresponding average predictor values. The DFBeta is a way to measure how much an observation has effected the estimation of a coefficient. It is very important to do the residual analysis.

Analysis:

The goal of our project is to find which variables influence the probability that an individual would pick up the nickel. We watched all videos, and used our collection rule to analysis which individual should be considered as valid data. Our variables are, gender, age, race, suit, doing something, empty hand, backpack type, with friends, surround has stranger, running, direction (in or out), and which side he/she walks(left or right).

The first step is to choose a significant model. As we have 12 variables, we use forward and backward stepwise to select the significant variables and remove the insignificant variables. We mainly use Omnibus test, Cox & Snell R square, Nagelkerke R square, and Hosmer and Lemeshow test. And then we look at the significance of the variables. It seems that backward stepwise is better than forward stepwise.

Then we checked the collinearity of the remaining variables model. We use the stepwise method to check the collinearity. We checked VIF, tolerance, standard error, condition index, and eigenvalue. After that we got a 3 variables model.

After we got the model we use 50-50 and 80-20 cross validation to verify our model. And also check the collinearity of this model.

We also analyzed the residuals to ensure that their behavior was well within acceptable levels. We mainly use Cook’s distance, leverage, DFBeta, standardized residual. It all works well.

Results:

Our model is related to 3 variables, which are with friends (alone = 0, with friends = 1), doing something (doing nothing = 0, doing something = 1), and direction (in = 0, out = 1). The model is :

$$\text{Logit(odds)} = -2.846 + 3.006*\text{doingsth} - 1.483*\text{friends} - 1.303*\text{direction}$$

The Wald statistic is as in Table 1. The coefficient is significance.

	B	**S.E.**	**Wald**	**Sig.**
doingsth	3.006	0.452	44.281	0.000
friends	-1.483	0.238	38.864	0.000
direction	-1.303	0.210	38.376	0.000

Table 1 Wald Statistic

The classification table is as below. The overall accuracy is 88.4%.

Observed	**Predicted**		
	Not Pick	Pick	Percentage Correct
Not Pick	933	30	96.9%
Pick	96	27	22.0%
Overall percentage			88.4%

Table 2 Classification Table

The Cox & Snell R square, Nagelkerke R square, and Hosmer and Lemeshow test is in table 2. We can see that the significance of Hosmer and Lemeshow test is 0.917. This means it is pretty good. But the Cox & Snell R square is 0.120 and Nagelkerke R square is 0.237, which are low.

Test	**Statistic**	**Sig.**
Hosmer and Lemeshow Test	0.952	0.917
-2 log likelihood	628.476	--
Cox & Snell R^2	0.120	--
Nagelkerke R^2	0.237	--

Table 3 Model Summary

We also looked at the correlation matrix to see whether it exists multicollinearity (Table 4). From the table we can see that the 3 variables has no multicollinearity. But we also checked eigenvalue, condition index, tolerance and VIF. The eigenvalues are all larger than 0. The condition indexs are all smaller than 10. The VIF are all smaller than 5. So we can see that these 3 variables has no multicollinearity.

	Constant	**doingSth**	**friends**	**direction**
Constant	1.000	-0.875	-0.070	-0.166
dotingSth	-0.875	1.000	-0.304	-0.028
friends	-0.070	-0.304	1.000	-0.068

direction	-0.166	-0.028	-0.068	1.000

Table 4 Correlation Matrix

	Eigenvalue	Condition Index	Tolerance	VIF
doingSth	0.766	1.790	0.805	1.242
friends	0.432	2.382	0.793	1.261
direction	0.348	2.654	0.982	1.018

Table 5 Multicollinearity

As our R square is small, I should make sure that the variables are significant in the model. So we did the 80/20 cross validation. We randomly split the data into to two parts, and check the model. We repeated this procedure 10 times. The result is as table 6. From the table we can see that the highest number is 0.017, which means our model is significant.

Iteration	Constant	doingSth	friends	direction
1	0.000	0.000	0.000	0.001
2	0.000	0.000	0.000	0.003
3	0.000	0.000	0.000	0.003
4	0.000	0.000	0.017	0.000
5	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000
9	0.000	0.000	0.001	0.000
10	0.000	0.000	0.000	0.000

Table 6 Cross Validation

We also analyzed the residual(Table 7). From the table we can see that Cook's distance is 0.2166 which is smaller than 1. The leverage value is 0.01204. According to the (k+1)/n=(4/1086)=0.004, so 0.01204 is 3 times of 0.004. It means this is acceptable. The DFBeta are all smaller than 1. What's more, we also analyzed the standard residual. The results are 96.8% in [-1.96, 1.96], 99.8% in [-2.58, 2.58], 100% in [-3, 3]. These all means that no case exerts an undue influence on the model

Test	**Maximum Value**
Cook′ s Distance	0.2166
Leverage Value	0.01204
DFBETA for Constant	0.17103
DFBETA for doingSth	0.02707
DFBETA for friends	0.04902
DFBETA for direction	0.02909

Table 7 Residual Analysis

Conclusion

We are able to use logistic regression to build a significant model to predict whether an individual will pick up the nickel when he/she saw it. The direction, whether with friends, and whether he/she is doing something are the important variables. If a person is with friends, doing nothing, and leaving the library are more possible to pick up the nickel.

The weaknesses are:

1. Some variables are subjective, such as age, race. Sometimes it is difficult to decide the values of these variables.
2. Maybe the front door of Doheny is not a good place to put the nickel, because only less than 10% people could notice the nickel.
3. As a few people could notice the nickel, it is not easy to collect data. We need more time and more data.

Reference:

[1] Christopher M. Bishop (2006). Pattern Recognition and Machine Learning. Springer. p. 205. "In the terminology of statistics, this model is known as logistic regression, although it should be emphasized that this is a model for classification rather than regression."

[2] Field, Andy. Discovering statistics using SPSS. Sage publications, 2009.