FEDERAL RESERVE BANK *of* ATLANTA WORKING PAPER SERIES

Alternative Methods for Studying Consumer Payment Choice

Oz Shy

Working Paper 2020-8
June 2020

Abstract: The study of consumer payment choice at the point of sale involves a classification of payment methods such as cash, credit cards, debit cards, prepaid cards, paper checks, and electronic payments withdrawn from consumers' bank accounts. I describe alternative methods for studying consumer payment choice using some machine learning techniques applied to consumer diary survey data. I then compare the results to the more traditional logistic regression methods. Machine learning techniques have advantages in generating predictions of payment choice, in visualization of the results, and when applied to high-dimensional data. The logistic regression approach has an advantage in interpreting the probability that a buyer uses a specific payment instrument.

JEL classification: C19, E42

Key words: studying consumer payment choice, point of sale, statistical learning, machine learning

https://doi.org/10.29338/wp2020-08

The author thanks participants at the Bank of Canada Retail Payments Workshop held in Ottawa October 24–26, 2018, for valuable comments on an earlier draft. The views expressed here are those of the author and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the author's responsibility.

Please address questions regarding content to Oz Shy, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree St. NE, Atlanta, GA 30309, Oz.Shy@atl.frb.org.

Alternative Methods for Studying Consumer Payment Choice*

Oz Shy†
Federal Reserve Bank of Atlanta

June 16, 2020

Abstract

The study of consumer payment choice at the point of sale involves a classification of payment methods such as cash, credit cards, debit cards, prepaid cards, paper checks, and electronic payments withdrawn from consumers' bank account. I describe alternative methods for studying consumer payment choice using some machine learning techniques applied to consumer diary survey data. The results are then compared to the more traditional logistic regression methods. Machine learning techniques have advantages in generating predictions of payment choice, in visualization of the results, and when applied to high-dimensional data. The logistic regression approach has an advantage in interpreting the probability that a buyer uses a specific payment instrument.

Keywords: Studying consumer payment choice, point of sale, statistical learning, machine learning.

JEL Classification Numbers: C19, E42.

(Draft = diary-ml-29.tex 2020/06/16 09:15)

*I thank participants at the Bank of Canada Retail Payments Workshop held in Ottawa October 24–26, 2018 for valuable comments on an earlier draft. Declaration of interest: None. The views expressed here are the author's and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

†*Email*: `Oz.Shy@atl.frb.org`. Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree St. NE, Atlanta, GA 30309, U.S.A.

1. Introduction

Data on "how consumers pay" are collected by consumer surveys in which consumers list all the payment instruments they have and whether and how they use them at the point of sale (POS). There are two types of consumer survey: Surveys that rely on memory recall (how many credit card payments have you made during the last year or month?) and *diary* surveys in which consumers record, either in real time or by the end of each day, all payment related activities including dollar amount, spending type, merchant type, and payment method. For budgetary reasons, consumer diaries (used in this paper) are limited to few thousands respondents thereby generating approximately between fifteen to thirty thousand payment observations. Whereas these numbers are considered to be "small data" relative to payment data collected by Visa, MasterCard, Amazon, and other retailers that collect "big data," consumer diary surveys are more suitable for analyzing consumer payment choice because they provide on a wide variety of spending types using *all* payment methods rather than just credit and debit card payments.

The study of consumer payment choice at the point of sale involves a classification of payment methods such as cash, credit cards, debit cards, prepaid cards, paper checks, and electronic payments withdrawn from consumers' bank account. Because these studies focus on *classification* of payment methods, researchers have traditionally relied on discrete-choice analyses using binomial, multinomial, logit, and probit regression methods. These studies measure mainly the effects of dollar transaction value (amount), demographic variables, and spending/merchant type on the probability that a consumer chooses to pay with a particular payment method such as credit card instead of cash.

Whereas standard regression models will remain a powerful tool for studying payment choice, the goal of this article is to explore alternative and somewhat newer methods involving machine (statistical) learning techniques. Machine learning methods introduce new insights to our attempts to predict how a specific consumer pays for a specific transaction at a particular POS, and these techniques become essential when using "big data." The recent adoption and uses of machine learning techniques in economics, marketing, and sales have been discussed in Varian (2014), Mullainathan and Spiess (2017), and references therein. The statistical properties of these techniques are discussed over the Internet and in several textbooks, such as James et al. (2013).

The list below provides a description of the payment methods collected by the data that are analyzed in this article and described in section 2. These descriptions have been provided to all the respondents who participated in the survey of payment choice.

Electronic payment methods:

- Bank account number payment (BANP): You [the respondent] give your bank routing number and account number to a third party to authorize a deduction from your bank account.
- Online banking bill pay (OBBP): You initiate a payment at your bank's online banking website.
- Account-to-account money transfers.
- Income deduction: Your employer makes the payment for you on your behalf and deducts the payment amount from your salary.

Card payment methods:

- Debit card: Your payments are deducted from your bank account. Also, you can use a debit card to withdraw or deposit cash at ATMs.
- Credit card: You pay back the credit card company later. Credit cards charge interest (unless the balance is paid by the initial due date).
- Prepaid/gift/EBT card: You store or load money on a prepaid card. Sometimes called "prepaid debit," "gift cards," "payroll cards," or "stored value cards."

Paper payment methods:

- Cash: Coins and paper bills.
- Check: You write a paper check to a person or business.
- Money order: You purchase a money order from a bank, post office, check-cashing store, or retail store. At the time of purchase, you specify the amount and the person or business to be paid.

This article is organized as follows. Section 2 describes the data and defines the variables of interest. Section 3 presents classification trees and shows how they can be used to visualize and predict consumer payment choice. The predictions are then compared to predictions generated by multinomial logistic regressions. Section 4 extends the classification tree analysis to random forest. Section 5 introduces the *k*-nearest neighbors algorithm. Section 6 illustrates the support vector machine algorithm. Section 7 concludes.

2. Data, variable selection, sample statistics, and coding

The data are taken from the 2019 Diary of Consumer Payment Choice (DCPC).[1] The DCPC uses a representative sample of U.S. (18 and older) consumers. The DCPC records transactions during three consecutive days. Transactions include purchases, bill payments, ATM withdrawals, and deposits. Respondents' three-day diaries were evenly distributed throughout the month of October 2019 in a way that resembles a three-period overlapping generations model. The DCPC collects a large number of variables describing all sorts of demographics and transactions. For the purpose of this article, I focus on only a small subset of variables.

Table 1 shows that out the 11,571 total payments made by 2,474 respondents, almost three-quarters (74.1 percent) were paid "in-person" and about one-quarter (25.9 percent) were paid remotely. Cash payments constitute 35.5 percent of in-person payments, followed by debit card payments (30.1 percent) and credit card payments (27.4 percent). In contrast, remote payments are more evenly distributed among debit cards (22.1 percent), followed by BANP (22.2 percent), credit card (18.2 percent) and paper checks (12.1 percent).

A note about terminology. In regression analysis, the *explanatory variables* are also called *independent variables*, *explanatory* variables, *regressors*, or *right-hand side variables*. In machine learning, the same variables are often called *features*, *attributes*, *predictors*, or *input variables*. The reader should bear in mind that the analysis in this paper uses all these terms to mean the same thing.

[1]The Federal Reserve Banks of Atlanta, Boston, and San Francisco conduct the diary. The results are summarized in Greene and Stavins (2019) and Kumar and O'Brien (2019). Similar surveys are conducted by the Bank of Canada; see Henry, Huynh, and Welte (2018). The data and assisting documents (codebooks) are publicly available for downloading from the Federal Reserve Bank of Atlanta website: https://www.frbatlanta.org/banking-and-payments/consumer-payments.aspx. The data and the R-code used in this analysis are available for downloading from the author's web page: www.ozshy.com (click on "Recent articles").

The remainder of this section describes eight features.[2]

- Amount: Numeric payment amount ranging from \$0.07 to \$29,000, with median \$25.00 and mean \$110.83.
- In-person: a categorical variables indicating whether the respondent paid in-person (8,570 payments or 74.1 percent) or remotely (3,001 payments or 25.9 percent). Remote payments are generally made online, deduction from a bank account, or via the mail.
- Age: Integer-valued ranging from 18 to 101, with median 55 and mean 53.37.
- Gender: 1,351 (54.6 percent) female respondents , 1,123 (45.4 percent) male respondents.
- Marital: 1,549 (62.6 percent) married, 925 (37.4 percent) not married.
- Education: 89 (3.6 percent) have elementary or less, 1189 (48.1 percent) have high school, 747 (30.2 percent) have an associate or a college degree, and 449 (18.1 percent) have MA or higher.
- HH_income: Household income ranging from \$0 to \$1m with median \$65,000 and mean \$78,665.
- HH_size: Number of people in the household ranging from 1 to 12 with median 2 and mean 2.65.

The analysis in this article does not apply demographic sampling weights because the whole purpose of this demonstration is to analyze the effects of demographics on payment choice (among other variables). In addition, because the goal is to compare very different statistical methods using the same data, I made sure that all these comparisons are conducted with the same training and testing subsamples which are sampled using the same pseudo-random number generating function (the *set.seed(N)* command in R).

3. Classification trees

In the context of machine learning, a classification tree displays an optimized algorithm in the form of an upside-down tree. The tree illustrates how the machine (software) splits and classifies

[2] I consolidated the number of levels within some categorical variables. For examples, the number "Education" levels was reduced from 16 to 4 which makes it easier to interpret the results. The number of "Marital" levels was reduced by combining divorced, separated, and widowed into "not married."

the data with the objective of minimizing a function of the number of classification errors among the predicted payment methods relative to the actually used payment methods.

The data set is split into two separate subsamples: a training subsample and a testing subasmple. The classification tree algorithm is designed and tuned on the training subsample using cross validation. The cross validation procedure partitions the training subsample into k folds, where the algorithm is constructed using $k-1$ folds of data and tested on the retained k's fold on which the classification errors are measured. The process repeats itself k times, each with a different retained k's fold. The k error measurements are then averaged to produce the final tree algorithm. The advantage of this method is that all observations in the training subsample are used for both training and validation.

The resulting tree algorithm is then evaluated by measuring the classification errors generated by predictions made on the testing subsample which was held out separately from the training subsample and the cross validation procedure. That is, testing data constitute a subsample from the same population but with observations that were not included in the sample used for fitting the model. Therefore, the predictive power of a model is assessed on a subsample of the original data set which is different from the training subsample that is used for the construction of a classification tree; a practice sometimes referred to as the "firewall principle."

To be able to compare the predictive power of the tree algorithm to that of logistic regressions, I duplicate the procedure by estimating the regression coefficients using the same training subsample that is used for designing the tree, and then measure the classification errors of the logistic regression model on the same testing subsample as the tree algorithm. Therefore, the sample with 11,571 payment observations described in section 2 is randomly split so that 9,256 (80 percent) of the payment observations are classified as training data, and the remaining 2,315 (20 percent) as testing data. The same split is retained throughout the analysis in order to be able to compare the efficiency (error minimization) of the various algorithms that are analyzed in this article.

In addition, the classification tree algorithm and the logistic regression are compared using exactly the same payment method classifications and the same explanatory variables that were described in section 2. The relationship between the choice of payment methods and the explanatory variables is defined by the model

$$\begin{aligned}\text{Method}_t = \alpha_m &+ \beta_T \text{Amount}_t + \beta_P \text{in-person}_t + \beta_A \text{Age}_t + \beta_G \text{Gender}_t \\ &+ \beta_M \text{Marital}_t + \beta_E \text{Education}_t + \beta_I \text{HH\_income}_t + \beta_S \text{Household\_size}_t,\end{aligned} \tag{1}$$

where the dependent variable "Method" is the payment method used for transaction (payment) t in the subsample. The demographic variables reflect those of the respondent who paid for transaction t.

3.1 Multiple payment methods classification tree

Figure 1 displays an upside-down tree that classifies the use of eight payment methods according to the model defined in (1). For the purpose of this illustration, the tree drawn in Figure 1 has been excessively pruned compared with the optimal tree size. That is, the displayed tree has been pruned to have only 5 splits whereas the tree that minimizes the average cross-validated error (plus one standard deviation) should have about 27 splits, which are hard to fit into a single page.[3]

Figure 1 shows that that the "in-person" feature is the most influential predictor for separating BANP from cash payments. This is because cash must paid in person whereas BANP is initiated electronically and remotely by the payee (such as a utility company). Note that the tree algorithm selects the "In-person" feature as the top split because this feature reduces classification errors more than a top split according dollar amount, age, or income (that are pushed to lower branches of the tree).

The second layer of branches in Figure 1 consists of splits according the payment dollar amount. For in-person payments, Figure 1 shows that a payment amount lower than \$20 constitutes the next-best predictor for paying with cash, which explains 40-percent of all respondents' payment observations. The exact splitting amount (\$20 in this sample) is determined by the error "majority rule" or some function of it in the sense that any other split would generate a lower prediction

[3] The classification tree displayed in Figure 1 was constructed with the *rpart* R package. Initially, the complexity parameter was set to $cp = 0.001$, thereby generating 41 splits. For a given *cp* value, the program attempts only the spits that increase the overall fit by a factor of *cp* or higher. Then, in order to increase prediction accuracy, the complexity parameter was increased to $cp = 0.001388889$ thereby reducing the number of splits to 27.

accuracy as measured by a function of the number of prediction errors.[4] For remote payments, large payment amounts (over $89) are predicted to be paid via BANP.

The third layer of branches in Figure 1 consists of splits according to education (for in-person payments) and household income for remote payments. These features predict whether the payment will be made with a credit card or a debit card. Note that high household income may be correlated with payers' ability to obtain and use credit cards, and this explains why high household income (not less than $157,000) predicts credit card payments for remote payments. For in-person payments, associate degree or college degree and above also predict credit card payments, most likely because these are correlated with household income.

The analysis of classification trees has so far focused on the visualization aspect. The analysis below demonstrates the predictive aspect of the classification tree algorithm. I first construct a longer-than-optimal classification tree using cross validations restricted to the training subsample for the purpose of determining and tuning the optimal number of splits. Then, the tree is pruned by increasing the value of the *cp* parameter which reduces the number of splits to the level that minimizes the cross-validation error.

Note that the reason why trees are pruned is to avoid the consequences of "overfitting." That is, longer trees reduce prediction error rates on the training data, but increase the prediction error rates when used over the testing data, and therefore result in poor prediction power over new data. To see that, suppose that we draw a very long tree in which each observation results in its own terminal node. This tree achieves the maximum possible fit over the training data, but will surely fail to provide proper predictions using the testing data. For this reason, the long tree should be pruned according to the desired *cp* parameter. The pruned tree is then used to make predictions using the testing data or any other new data from the same population.

Table 2 exhibits the *confusion matrix* (also called error matrix). The values on the main diagonal are the numbers of correct predictions using the testing data. That is, the optimally-tuned tree correctly predicted 58 BANP payments, 0 Acct2acct payments, 396 cash payments, 5 check pay-

[4]The reader may wonder why the tree algorithm selects the $20 payment amount as the optimal split (and not $19 or $21). This may have to do with the dollar denomination structure which causes consumers to carry $5, $10, and $20 bills in their pocket. The $20 split can also be explained by the observation that most ATMs in the U.S. dispense only $20 bills. Discontinuities of this type have been analyzed in Shy (2020) using U.S. diary data.

ments, 268 credit card payments, 307 debit card payments, 44 OBBP and 0 prepaid card payments. The failure to predict any Acct2acct and prepaid card payments stems from the fact that splitting features according to the level of use of payment methods did not yield sufficient reduction in the error rates. This is because the data contain only a small number of Acct2acct and prepaid card payment observations.

The off-diagonal values reflect incorrect predictions. For example, 84 predictions that consumers pay with credit cards and 120 that consumers pay with debit cards were incorrect because these consumers actually paid cash.

The bottom row in Table 2 computes the tree algorithm's correct prediction rate for each payment method. It shows that out 602 actual cash payments in the testing data, the tree was able to accurately predict 66 percent. Out of 571 credit card payments in the testing data, the tree predicted only 47 percent accurately. Debit card payments were correctly predicted 47 percent, and as mentioned above, there were zero correct predictions of Acct2acct and prepaid card payments.

3.2 A comparison with multinomial regressions

Multinomial logit regressions are commonly used in payments research. Therefore, I now compare the prediction accuracy of a multinomial logit regression to the prediction accuracy of the tree algorithm, using the exact same model defined in equation (1).

The regression is run on exactly the same training subsample of 9,256 payments as the classification tree. The predictions are then performed on the same testing subsample with the same 2,315 payments as the classification tree. This is needed in order to be able to compare the number of prediction errors between the two algorithms.

The table of coefficients of a multinomial regression is too long to be displayed here. Instead, Table 3 presents the confusion matrix and the prediction accuracy rate for the multinomial logit model.[5] Comparing Table 3 with Table 2 shows that the classification tree algorithm has higher prediction accuracy rates for BANP (43 versus 32 percent), Cash (66 versus 63 percent), and debit card payments (47 versus 43 percent). However, the classification tree has slightly lower accurate prediction rates of Checks (3 versus 11 percent) and OBBP (36 versus 37 percent).

[5]Results were obtained using the *multinom* function in the *nnet* R package.

Therefore, we cannot conclude which algorithm predicts better at least for this sample. Classification trees have the advantage that they generate intuitive graphical presentations of payment method classifications, such as the one shown in Figure 1. However, regression coefficients (especially marginal effects) are often easier to interpret.

4. Random forest

Random forest algorithms do not generate classification trees like the one plotted in Figure 1. Instead, they build a large number of classification trees where each tree splits the training data according to a randomly-selected subgroup of features on each bootstrapped sample. The number of features in each subgroup is controlled by the researcher, and is generally chosen to be near the square root of the total number of features. More specifically, the specification of the model in equation (1) analyzes the data with the use of eight features. Therefore, each tree in the forest will split the data according to three randomly-selected features. This procedure decorrelates the sampled trees in the sense that strong features (such as "in-person" and"Amount" in our data), will not be chosen as top splits in every tree.

Although random forest algorithms do not yield trees that can be drawn, they provide an important procedure for variable selection (in addition to improved prediction accuracy). Random forest graphs rank the relative importance of the features for prediction purposes. Figure 2 plots a variable importance graph using R's *randomforest* package. It shows that removing the predictor "In-person" from the data would increase the average number of misclassifications by 219. Removing the predictor "Amount" from the sample would increase the average number of misclassifications by 198. As the figure shows, "Marital" status is the least important variable because without it the average number of misclassifications would increase by 67 only.[6]

Most importantly, random forests are extremely useful for making predictions. The model defined in (1) is now applied to the random forest algorithm. The algorithm uses exactly the the same training subsample containing 80-percent of the observations that were used to fit the classification tree and the multinomial regression that were analyzed in section 3. The algorithm's

[6]Figure 2 provides an easy visual framework for variable selection. Other commonly used variable selection methods include Lasso and Ridge regressions and principal component analysis that can be applied via several R packages.

prediction accuracy is then tested on the same remaining 20-percent of the data. The resulting confusion matrix for the random forest predictions is given in Table 4.

Comparing Table 4 with Tables 2 and 3 reveals significant improvements in prediction accuracy for all payment methods. Perhaps the most striking result is that, unlike the tree algorithm and multinomial logit, random forest is the only prediction algorithm that generated some accurate predictions for account-to-account transfers and prepaid card payments. More precisely, random forest predicted correctly 2 out of 28 account-to-account transfers and 6 out of 44 prepaid card payments whereas the classification tree and multinomial logit predicted none. Although these prediction rates may seem to be low, the reader should bear in mind that they were made on the testing subsample which contains totally different observations than the training subsample that was used to fit all the classification trees in the forest.

The reason for the significant improvement in prediction accuracy is that the random forest algorithm averages the predictions made by hundreds of trees, where each tree is constructed based on only three out of eight features. Therefore, dominant features, such as "In-person" and "Amount", are not always used, thereby letting a large number of trees be split based by less-dominant features. This is in addition to bootstrapping used for each tree prediction. This exercise shows, perhaps paradoxically, that introducing more randomness into the prediction algorithm tends to improve prediction accuracy.

5. k-nearest neighbors (kNN)

The k-nearest neighbors is an algorithm that inputs available observations and classifies new observations based on a similarity measure in the space of features. Similarity is determined by the distance between the features of the observations. The kNN algorithm determines a neighborhood in the space of features for each observation in the training data. Each neighborhood contains k observations, where k serves as a tuning parameter to be determined by the researcher.

Predictions of which payment method will be used are based on measuring the distance in the space of features to the k nearest neighbors, and determining the payment method associated with the majority of these k nearest observations. For example, if $k = 3$, the algorithm measures the distance between the features of the new observation to its $k = 3$ closest neighbors. If two of the its neighbors (say from the training subsample) are credit card payments and one is a cash payment, the new observation will be classified as a credit card payment by a simple majority rule.[7]

One drawback of the kNN algorithm is that it can analyze only *numerical* features because neighborhoods are defined according to distances among observations within each neighborhood (such as the Euclidean distance). This is unfortunate because most of the data collected in payments research contain categorical features such as transaction type and payers' demographic variables. A partial solution to this drawback would be to turn categorical features into numerical. Therefore, for the *k*NN analysis, "In-person," "Married," and "Male" are assigned the value of 1 (versus 0), and the four levels of "Education" are assigned the values of 0, $\frac{1}{3}$, $\frac{2}{3}$ and 1.

Another potential problem that may arise from the fact that neighborhoods rely on distance measurements is that changing the units in which features are measured may change the neighborhoods determined by this algorithm. For this reason, all the features are normalized to take values between 0 and 1 before running the kNN algorithm. Formally, each numerical feature, for example, a particular household income $\hat{x}$ is normalized according to $[\hat{x} - \min(x)]/[\max(x) - \min(x)]$, for each household income $\hat{x}$ in the sample.

5.1 Illustration of the *k*NN algorithm using two features

The grouping of training payment observations into neighborhoods in the space of features cannot be plotted in two dimensions when there are more than two features. This is unfortunate because the algorithm is suitable for making predictions using all the eight features that are listed in the model defined in (1). Therefore, in order to illustrate these neighborhoods, I focus only on two numerical features: payment dollar amount (Amount) and household income (HH_income). These two features are normalized to take values between 0 and 1 to avoid the influence of units of measurements.

In order to avoid dealing with the categorical "In-person" feature which can take only two

[7] To some degree, kNN resembles clustering analysis (k-means) which partitions the space of features into clusters according to distance (such as the Euclidean distance). Clustering analysis is useful for *unsupervised* data which contain only features and no response variables. The data analyzed in this article is *supervised* because each observation has a response variable which is choice of payment method.

values, I also restrict this illustration to 8,570 in-person payments. This subsample of in-person payments is then divided into 6,856 (80 percent) training observations and 1,714 (20 percent) testing observations.

The top part of Figure 3 depicts the actual 1,714 payment observations in the test data. To avoid cluttering, the axes are stretched by taking the logarithm of the normalized household income and payment amounts. The top part shows that consumers with lower household income use cash more intensively than consumers with higher income. Cash is also used more often for low dollar amounts. Checks are used for large dollar amounts. Credit cards are used more intensively by consumers with higher household income.

The bottom part of Figure 3 depicts 1,714 payment methods that were predicted from the two features (Amount and HH_income) in the test data, based on the neighborhoods determined by the *k*NN algorithm using the training subsample. In this exercise, the optimal (error-reducing) neighborhood size was found to be $k = 34$ observations. The bottom part shows that the *k*NN algorithm predicts a higher number of cash payments made by respondents with lower household income relative to the actual cash payments (depicted on the top part). It also predicts less cash payments by respondents with very high household income. In addition, the algorithm predicts slightly more intensive use of checks for higher dollar amounts relative to actual check payments.

5.2 *k*NN's predictions of payment methods based on all eight features

I now apply the *k*NN algorithm to the same training subsample containing 9,256 (80-percent) observations and test its prediction accuracy on the remaining 2,315 (20-percent) observations.[8]

Table 5 displays the confusion matrix for the payment method predictions made by the *k*NN algorithm applied to the same testing data used for the other three algorithms that were analyzed in previous sections. Comparing Table 5 with Table 4 reveals that, with the exception of prepaid card and check payments, the *k*NN predictions are less accurate than the random forest

[8]This analysis applies the *knn* function in the *class* R package. This function integrates the training data classification (first step) and the prediction on the testing data (second step) into a single function. The parameter k (number of observations in each neighborhood) is tuned by running a loop over this function for all values of $k = 1, \dots, 40$, yielding $k^{\min} = 3$ as the value that minimizes the classification error rate. Note that each time R runs this code may generate a slightly different value of $k^{\min}$ because the *k*NN classification algorithm is sensitive to the arbitrary choice of initial neighborhoods.

predictions. However, a comparison with Table 2 (classification tree) and Table 3 (multinomial regression) shows that the *k*NN algorithm generates more accurate predictions for all payment methods with the exception of BANP , cash, and OBBP. Moreover, just like the random forest algorithm, the *k*NN is exceptionally good in predicting payment methods that are less frequently used, such as account-to-account and prepaid card payments, which both the classification tree and multinomial regression algorithms failed to predict.

6. Demonstration of the support vector machine algorithm

Another commonly-used algorithm for analyzing supervised data is the support vector machine (SVM). The SVM algorithm is limited to a classification of two payment methods in the same way that a binomial regression is limited to a selection between only two options. Therefore, the SVM can be used to conduct research on cash users in order predict the use of cash versus non-cash payment methods. Similarly, the SVM can be used to research the use of card payments versus non-card payments. However, just like a binomial regression, there are no restrictions on the number of features (explanatory variables).

The SVM algorithm computes linear or nonlinear boundaries in the space of features. To illustrate the SVM, I simplify the regression model defined in (1) by splitting all the eight payment methods into two groups: cash payments versus non-cash payments. In addition, the number of features is restricted to "Amount" and "Age" because additional features cannot be drawn in two dimensions.

The black dots depicted in Figure 4 are 2,669 observed in-person cash payments. The dollar amount of each payment is given on the vertical axis. The age of the respondent who made the payment is given on the horizontal axis. Similarly, the red dots are 3,671 observed in-person non-cash payments.[9]

The grey-shaded areas in Figure 4 are generated by the SVM algorithm. These areas are the combinations of the respondents' age and dollar amounts that are predicted by the SVM algorithm to be paid with cash. The non-shaded areas are the respondents' age and dollar amounts that are

[9]The analysis in this section applies radial kernel option of the *svm* function in the *e1071* R-package.

predicted to be paid with non-cash payment instruments.

The shaded areas in Figure 4 imply that the SVM algorithm predicts low-value in-person payments to be made with cash and high dollar amounts to be paid with non-cash methods. In addition, in-person cash payments are more likely to be used by young consumers between the age of 18 and 28 and for payments below \$10 by consumers ages between 32 and 50. Figure 4 shows that the use of cash increases with age over 52.

Finally, the SVM algorithm predicts payments in the amount of \$20 to be paid with cash. These are illustrated by the shaded areas around \$20 dollar amounts drawn in Figure 4. Recall that the same result was predicted by the classification tree depicted in Figure 1. This happens because many ATMs in the U.S. dispense only \$20 bills. Therefore, cash users tend to carry multiples of \$20 bills in their wallets. Note also that there are shaded areas around the \$40 payment amounts that can be paid with two \$20 bills. Currency denomination also impacts this behavior as reflected by the shaded areas around the \$10 payment amounts and also around the \$25 and \$30 payment amounts. Footnote 4 provides a more detailed explanation for these observations.

7. Conclusion

Statistical (machine) learning algorithms provide useful tools for understanding and predicting consumer payment choice. This article samples a few of these tools using diary survey data on consumer payment choice.

The advantages of using these tools relative to conventional regressions are:

(a) They can be implemented on datasets containing a large number of explanatory variables (called features in statistical learning).

(b) Do not rely on restrictive assumptions on the structure of the data for ensuring consistency of the estimated parameters.

(c) Provide useful visualization tools of the results.

(d) Some algorithms, such as random forest, provide more accurate out-of-sample predictions.

(e) In particular, this article shows that both the random forest and the *k*NN algorithms are extremely useful in predicting consumers' choice of payment methods that are not used very often (such as account-to-account and prepaid card payments).

Some of the disadvantages of using these algorithms include:

(i) Loss of some interpretation power compared to the interpretability of estimated coefficients of linear regressions.

(ii) The tools are sensitive to their tuning parameters (regularization), to some degree similar to how regression results are sensitive to the choice of model, control of confounders, and the choice of interactions.

Anther commonly-used algorithm for analyzing supervised data, which is not demonstrated in this article, is neural networks. Neural network analysis attempts to mimic the way in which information transmission and decisions are made in the human brain.[10] Running neural network algorithms generally requires heavier computational power. In addition, neural network algorithms may encounter convergence problems for large number of classifications (such as eight payment methods) and features, such as attempting to predict eight payment methods using eight features.

To conclude, machine learning techniques and regression analyses could be viewed as complements rather than as substitutes. Using tools from both disciplines could only enhance the reliability of algorithms that predict consumer payment choice.

References

Greene, Claire and Joanna Stavins. 2019. "2018 Diary of Consumer Payment Choice." Federal Reserve Bank of Atlanta, Research Data Report No. 2019-3.

Henry, Christopher, Kim Huynh, and Angelika Welte. 2018. "2017 Methods-of-Payment Survey Report." Bank of Canada Staff Discussion Paper 2018–17.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*, vol. 112. Springer.

Kumar, Raynil and Shaun O'Brien. 2019. "2019 Findings from the Diary of Consumer Payment Choice." Federal Reserve Bank of San Francisco.

[10] Liébana-Cabanillas and Lara-Rubio (2017) investigate which merchant characteristics affect merchants' decision to adopt a mobile payment system by comparing a logistic regression model with a neural network analysis. The authors show that the neural network algorithm is the most precise tool in this research when predicting the use of mobile payment systems in certain lines of business.

Liébana-Cabanillas, Francisco and Juan Lara-Rubio. 2017. “Predictive and explanatory modeling regarding adoption of mobile payment systems.” *Technological Forecasting and Social Change* 120:32–40.

Mullainathan, Sendhil and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2):87–106.

Shy, Oz. 2020. “How Currency Denomination and the ATM Affect the Way We Pay.” *Journal of Economics and Business* .

Varian, Hal. 2014. “Big data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2):3–27.

	Electronic payments			Card payments			Paper methods		
Variable	BANP	OBBP	Acct2acct	Debit	Credit	Prepaid	Cash	Check	All
Number of payments	700	631	94	3247	2894	206	3082	717	11571
Percentage (%)	6.0	5.5	0.8	28.1	25.0	1.8	26.6	6.2	100.0
Avg. value ($)	349.94	428.53	408.21	51.30	66.17	32.73	23.97	404.53	110.83
Med. value ($)	144.60	120.00	250.00	25.00	29.97	12.00	10.00	100.00	25.00
Number of in-person	35	24	11	2583	2349	170	3045	353	8570
In-person (%)	0.4	0.3	0.1	30.1	27.4	2.0	35.5	4.1	74.1
In-person avg. value ($)	566.06	469.56	471.94	39.37	57.40	24.51	23.83	460.89	59.77
In-person med. value ($)	150.00	145.47	250.00	22.55	27.23	11.19	10.00	90.00	19.69
Number of remote	665	607	83	664	545	36	37	364	3001
Remote (%)	22.2	20.2	2.8	22.1	18.2	1.2	1.2	12.1	25.9
Remote avg. value ($)	338.56	426.91	399.76	97.70	103.99	71.54	35.14	349.88	256.66
Remote med. value ($)	143.87	120.00	250.00	44.66	44.90	17.23	20.00	100.00	80.00

Table 1: Sample statistics by payment method.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: The table displays information on 11,571 payments made by 2,474 respondents. Payment methods with less than 0.5% market share are excluded. Payment methods are described in section 1.

Payment Method	BANP	Acct2acct	Cash	Check	Credit	Debit	OBBP	Prepaid
BANP	58	15	0	23	20	38	40	3
Acct2acct	0	0	0	0	0	0	0	0
Cash	0	0	396	11	130	161	1	20
Check	2	1	1	5	4	3	0	0
Credit	21	3	84	50	268	136	19	4
Debit	18	4	120	35	125	307	19	17
OBBP	37	5	1	27	24	15	44	0
Prepaid	0	0	0	0	0	0	0	0
Total actual	136	28	602	151	571	660	123	44
Correct predictions	58	0	396	5	268	307	44	0
Correct rate (%)	43	0	66	3	47	47	36	0

Table 2: Confusion matrix for the classification tree.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 9,256 training payment observations and 2,315 testing payment observations. The main diagonal displays number of correct predictions.

Payment Method	BANP	Acct2acct	Cash	Check	Credit	Debit	OBBP	Prepaid
BANP	43	15	1	18	11	24	30	2
Acct2acct	0	0	0	0	0	0	0	0
Cash	0	0	379	24	150	233	0	17
Check	1	1	1	17	4	0	4	0
Credit	17	4	118	31	269	105	16	4
Debit	37	2	103	36	105	281	28	21
OBBP	38	6	0	25	32	17	45	0
Prepaid	0	0	0	0	0	0	0	0
Total actual	136	28	602	151	571	660	123	44
Correct predictions	43	0	379	17	269	281	45	0
Correct rate (%)	32	0	63	11	47	43	37	0

Table 3: Confusion matrix for the multinomial regression.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 9,256 training payment observations and 2,315 testing payment observations. The main diagonal displays number of correct predictions.

Payment Method	BANP	Acct2acct	Cash	Check	Credit	Debit	OBBP	Prepaid
BANP	56	12	2	19	17	26	32	2
Acct2acct	0	2	0	0	0	0	0	0
Cash	0	1	419	12	89	128	1	19
Check	6	2	1	32	15	6	16	0
Credit	17	5	85	39	363	78	10	4
Debit	27	2	94	32	64	406	10	13
OBBP	30	4	1	15	22	14	54	0
Prepaid	0	0	0	2	1	2	0	6
Total actual	136	28	602	151	571	660	123	44
Correct predictions	56	2	419	32	363	406	54	6
Correct rate (%)	41	7	70	21	64	62	44	14

Table 4: Confusion matrix for the random forest.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 9,256 training payment observations and 2,315 testing payment observations. The main diagonal displays number of correct predictions.

Payment Method	BANP	Acct2acct	Cash	Check	Credit	Debit	OBBP	Prepaid
BANP	48	10	2	12	35	19	26	3
Acct2acct	5	1	0	1	1	3	3	1
Cash	1	1	328	20	84	103	0	13
Check	11	1	9	31	21	25	13	1
Credit	17	3	111	33	330	91	24	6
Debit	27	6	140	38	75	394	13	12
OBBP	26	6	3	13	18	18	44	1
Prepaid	1	0	9	3	7	7	0	7
Total actual	136	28	602	151	571	660	123	44
Correct predictions	48	1	328	31	330	394	44	7
Correct rate (%)	35	4	54	21	58	60	36	16

Table 5: Confusion matrix for the *k*NN.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 9,256 training payment observations and 2,315 testing payment observations. The main diagonal displays number of correct predictions.

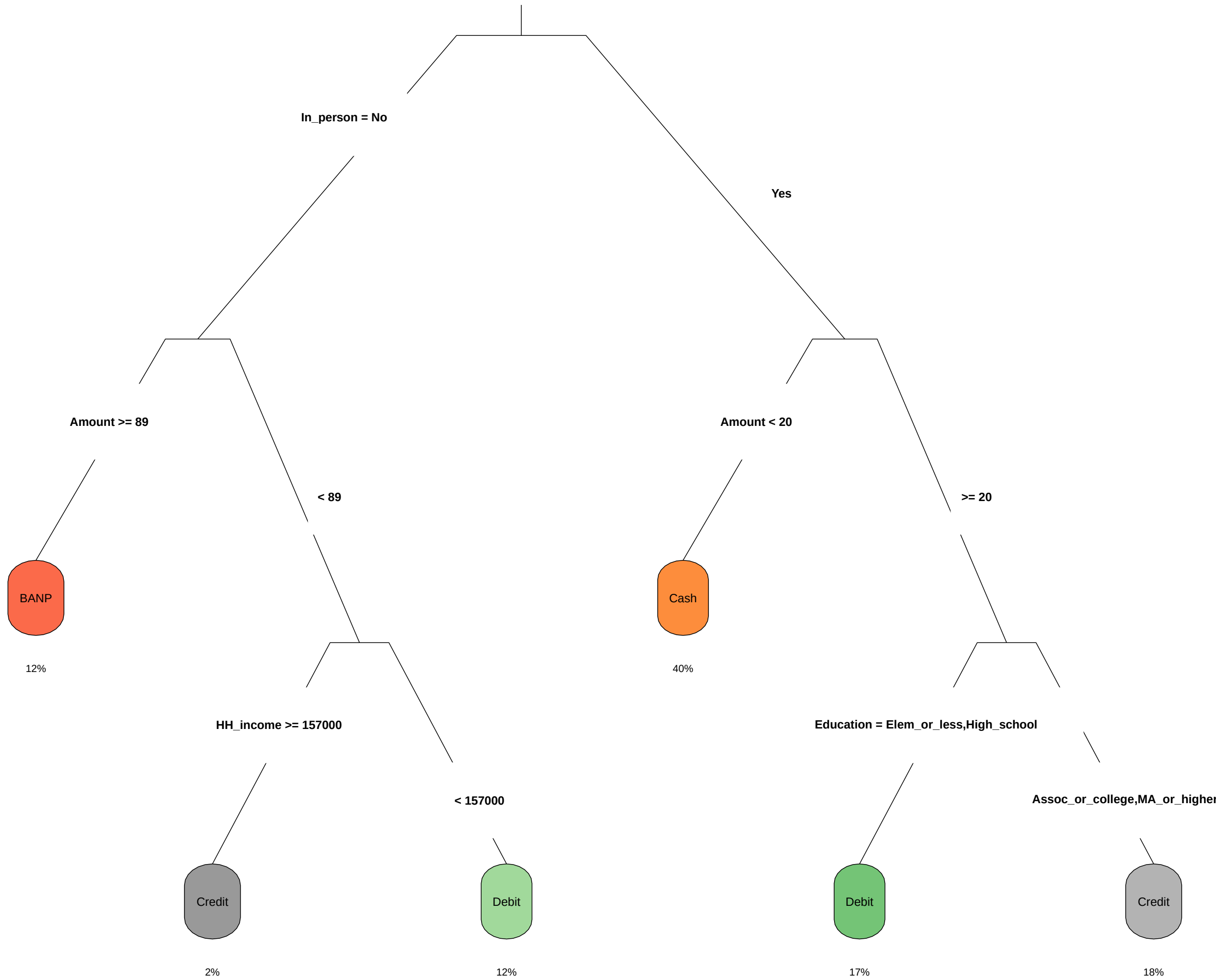

Figure 1: Classification tree: Predicting the use of payment methods based on whether the payment is made in-person, payment dollar amount, and some other consumer demographic variables.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 11,571 payment observations made by 2,474 respondents. For illustrative purposes, the displayed tree has been pruned to have only 5 splits instead of 26 splits which minimize the cross-validation errors.

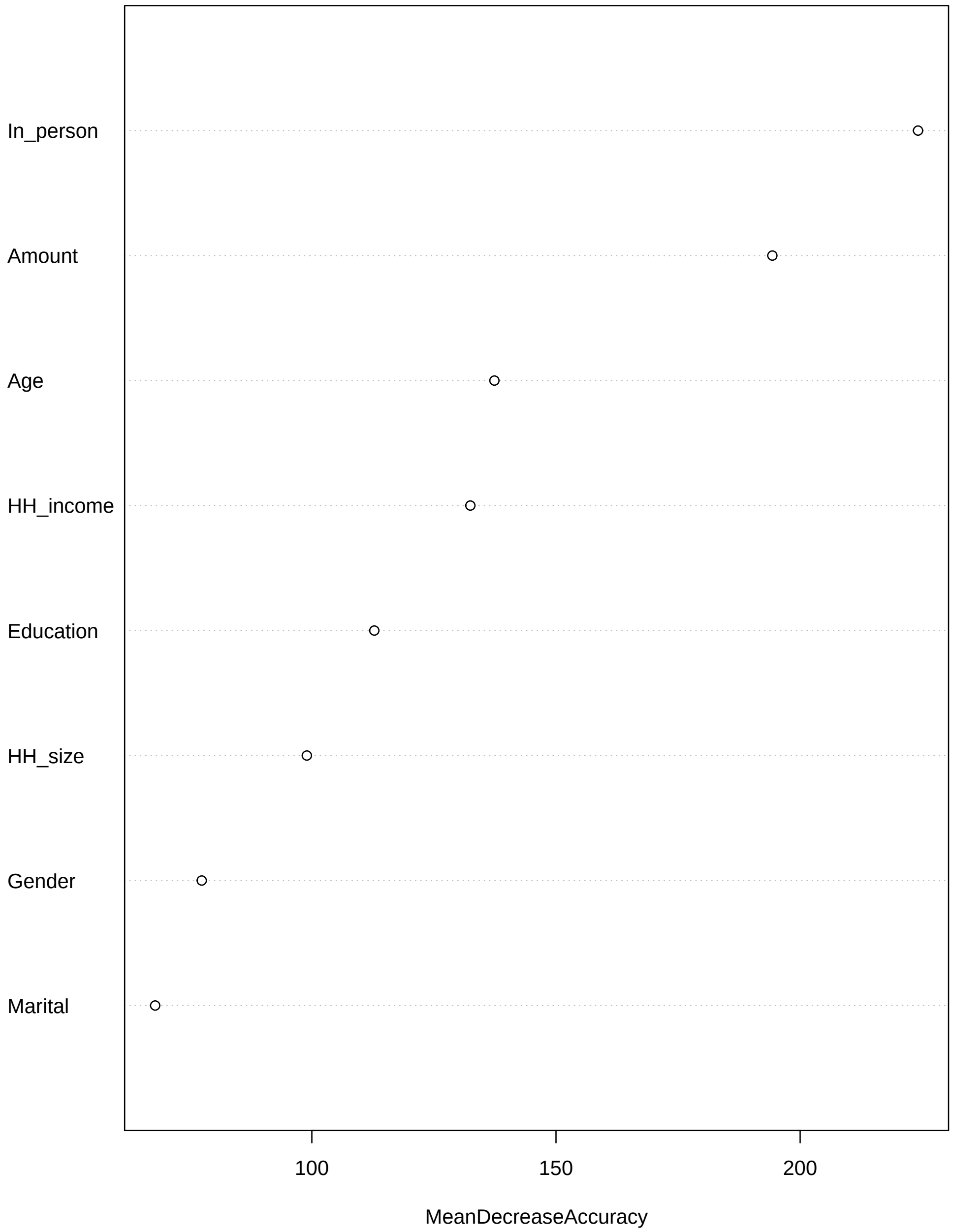

Figure 2: Variable importance plot (VIP) generated by random forest.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 9,256 training payment observations.

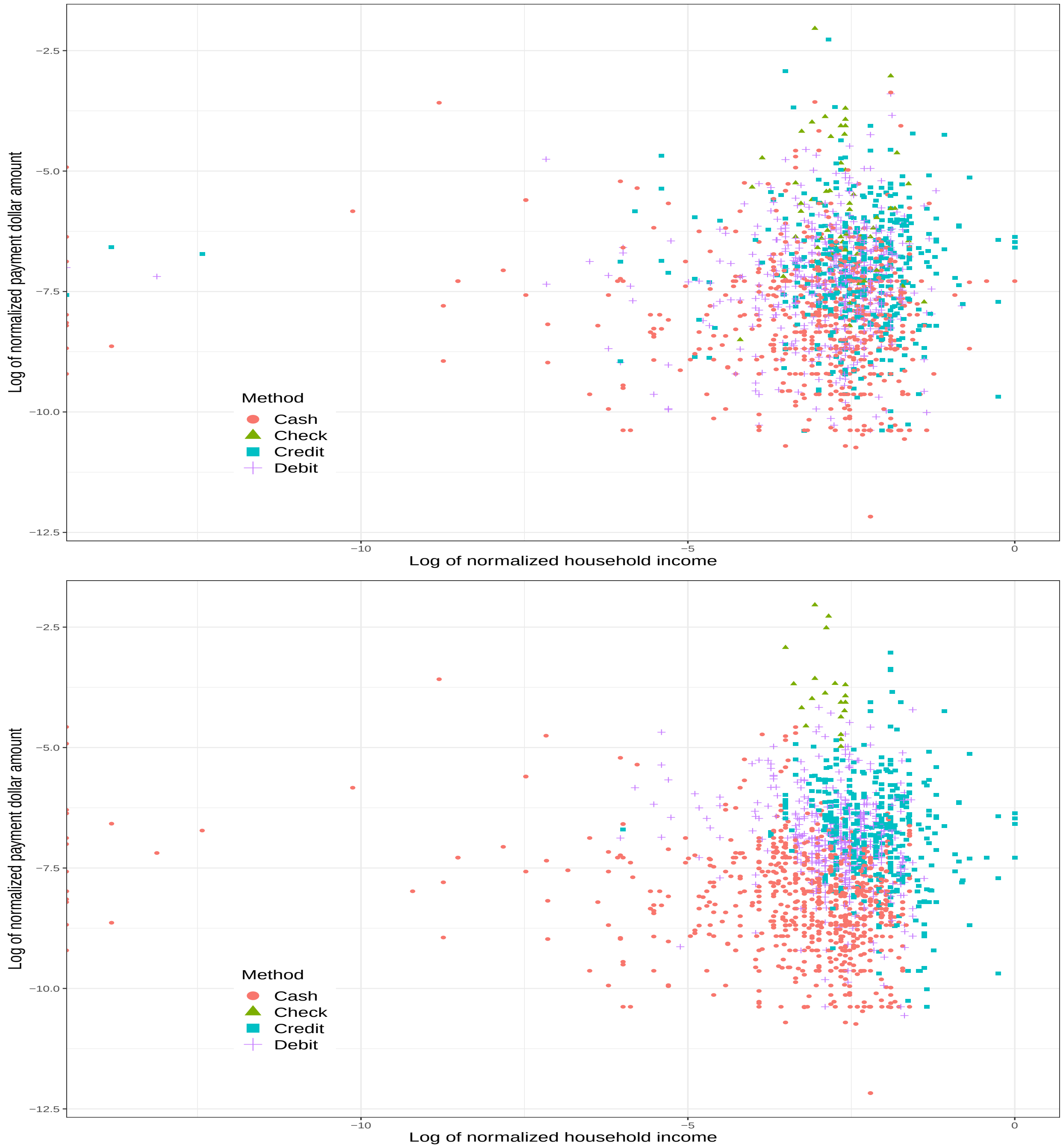

Figure 3: kNN classifications of in-person payment methods in the space of two features: Amount and household income.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 6,856 training in-person payment observations and 1714 testing in-person payment observations. *Top*: Actual in-person payment observations. *Bottom*: Predicted payment methods based on two features in the testing data.

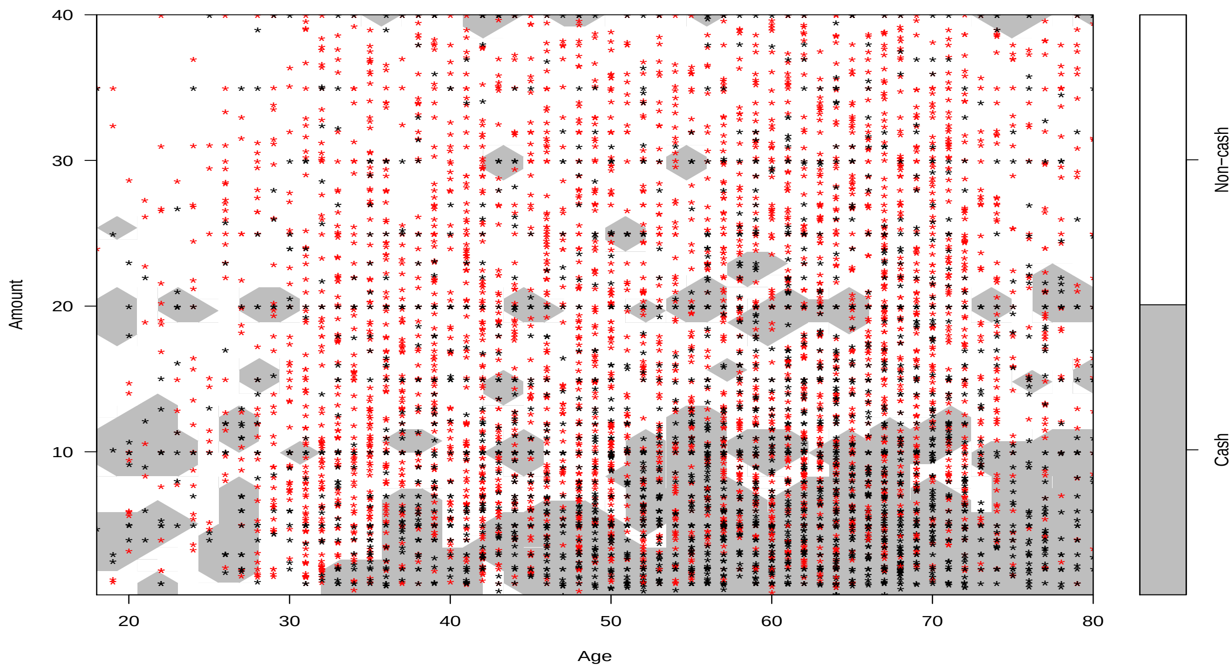

Figure 4: SVM payment method classifications in the space of two features: Amount and age.

Source: Author's computations from the 2019 Survey and Diary of Consumer Payment Choice.
Notes: Based on 6,340 (2,669 cash and 3,671 non-cash) in-person payment observations.