week4

Paul Anderson

9/13/2017

Rundown

The purpose of this week is to further develop your ideas for the competition and to explore the sample code provided in this markdown document. I have gone through your markdown documents from last week, and I the following list of general directions that people are taking:

- Budgeting related - my only concern with this is to make sure it is analytics or data science based. Adding a prediction/classification/clustering/ranking component using the data.
- Rating loyalty
- Dynamic links/recommendations on the web
- Financial goal - again make sure you think ranking/prediction/classification/etc
- Prob of taking/having a loan
- Predicting what purchases a person will make for incentive purposes
- Targetted rewards
- Predicting type of consumer

Loading the data

This is primarily from last week, but I'm including it here as well.

```
library(readxl)
month_end_balances <- read_excel("/usr/local/Learn2Mine-Main/galaxy-dist/lesson_datasets/Fake+Data+and+
    sheet = "Month end balances ", col_types = c("numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric"))

month_end_balances$mortgage_flag = factor(month_end_balances$mortgage_flag )

daily_interactions_WF <- read_excel("/usr/local/Learn2Mine-Main/galaxy-dist/lesson_datasets/Fake+Data+a
    sheet = "Daily interactions with WF")

daily_interactions_WF$Des1 = factor(daily_interactions_WF$Des1)
levels(daily_interactions_WF$Des1)
```

```
##   [1] "Account Closed"
##   [2] "ACCOUNT HISTORY COPY REQUEST"
##   [3] "Account Hold Add"
##   [4] "Account Hold Remove"
##   [5] "Account Inquiry"
##   [6] "Account Maintenance"
##   [7] "Account Open"
```

```
##   [8] "Account Open - IRA Account"
##   [9] "ACH Prenote"
##  [10] "Add Banker Note"
##  [11] "Add Contact Event"
##  [12] "ADDRESS CHANGE"
##  [13] "Advance"
##  [14] "Advance Reversal"
##  [15] "Agent Call"
##  [16] "ATM/CHECK CARD MAINTENANCE"
##  [17] "ATM/DEBIT PIN CARD"
##  [18] "ATM Failure"
##  [19] "ATM Time Out"
##  [20] "AUTHENTICATION_TRACKER_ON_OFF"
##  [21] "Authorized Credit"
##  [22] "Authorized Debit"
##  [23] "Balance Inquiry"
##  [24] "Balance Transfer Initiated"
##  [25] "BALANCE TRANSFERS"
##  [26] "Bank-initiated debit"
##  [27] "Bank-initiated transfer"
##  [28] "Bank Product Purchase"
##  [29] "Bank Product Purchase Reversal"
##  [30] "Bill Payment Miscellaneous"
##  [31] "Bill Payment Reject"
##  [32] "Bill Payment Reversal"
##  [33] "Book Transfer Create"
##  [34] "Business Credit Only flow selection"
##  [35] "Business Deposit and Credit flow selected"
##  [36] "Business Deposit - IOLTA flow selection"
##  [37] "Business - Deposit - RETA flow  selected"
##  [38] "Business  Deposit Only - Special Relationship flow"
##  [39] "Cancel Contact Event"
##  [40] "Cancelled"
##  [41] "Cash Check on Credit Card/LOC"
##  [42] "Cash EE Bond"
##  [43] "Cash Non-WFB Check"
##  [44] "Cash WFB Check"
##  [45] "Cash WFB Check-OWNER"
##  [46] "Check"
##  [47] "Check Card Credit"
##  [48] "Check Card Purchase Preauthorization"
##  [49] "Check Card Purchase Transaction"
##  [50] "CHECK ORDER"
##  [51] "CIVSALES_CARDS (PI)"
##  [52] "CIVSALES_CIP_UPDATE"
##  [53] "CIVSALES_CIP_VALIDATION"
##  [54] "CIVSALES_CREDIT_OPTION_GUIDE"
##  [55] "CIVSALES_CUST_NEEDS_ASSESSMENT"
##  [56] "CIVSALES_CUSTOMER_OFFERS"
##  [57] "CIVSALES_CUSTOMER_SESSION"
##  [58] "CIVSALES_CUST_PROFILE_EDITS"
##  [59] "CIVSALES_NEW_CUSTOMER_PROFILE"
##  [60] "CIVSALES_NEW_PMA"
##  [61] "CIVSALES_ONLINE_BANKING_BILL_PAY"
```

```
##  [62] "CIVSALES_PARTNER_REFERRAL_CREATED"
##  [63] "CIVSALES_PMA_ACCOUNT_CONVERSION"
##  [64] "CIVSALES_PMA_ADD/REMOVE_OWNERS"
##  [65] "CIVSALES_REPORT_REASON_FOR_CALL"
##  [66] "CIVSALES_RISK_SCREENING"
##  [67] "CIVSALES_SAVE_AS_YOU_GO_MAINTENANCE"
##  [68] "CIVSALES_TAB_CLICKER"
##  [69] "CIV_SPECIAL_RATES"
##  [70] "CIVSSALES_PMA_ACCOUNT_LINKAGES"
##  [71] "CLAIMS_PROCESSING"
##  [72] "CLIENT_SYSTEM_INFO"
##  [73] "Close"
##  [74] "CNA Added"
##  [75] "CNA Updated"
##  [76] "College Information Maintenance"
##  [77] "Common Customer Event History tool Account Level"
##  [78] "Common Customer Event History tool Customer Level"
##  [79] "Competitor Accounts Inquiry"
##  [80] "Complete Contact Event"
##  [81] "Consumer  Credit Only flow selected"
##  [82] "Consumer Deposit and Credit flow selected"
##  [83] "Consumer Deposit Only flow selected"
##  [84] "Consumer Establish/Maintain Remittance Agreement"
##  [85] "Consumer  New ATM/Check Card Only flow selected"
##  [86] "Consumer Recommendations Create MSR"
##  [87] "CONTACT_EVENT_CREATED"
##  [88] "Correction"
##  [89] "CORRESPONDENCE"
##  [90] "CreateMSR"
##  [91] "Credit Adjustment"
##  [92] "Credit Application"
##  [93] "CREDIT_CARD_FEE_REIMBURSEMENT"
##  [94] "CREDIT CARD REWARDS"
##  [95] "Credit Offer at ATM"
##  [96] "Credit Options guide (COG) tool accessed"
##  [97] "Credit Reversal"
##  [98] "CUAC Maintenance"
##  [99] "Customer Address Change"
## [100] "UpdateMSR"
## [101] "Verify Non-WFB Funds"
## [102] "Verify WFB Funds"
## [103] "Wire Transfer Create"
```

Some good ways to look at the data initially

```
colnames(month_end_balances)
```

```
##  [1] "masked_id"                "asof_yyyymm"
##  [3] "age"                      "tenure_altered"
##  [5] "checking_acct_ct"         "savings_acct_ct"
##  [7] "mortgage_flag"            "heloc_flag"
##  [9] "personal_loan_flag"       "cc_flag"
## [11] "prot_acct_flag"           "check_bal_altered"
```

```
## [13] "sav_bal_altered"           "mortgage_bal_altered"
## [15] "heloc_bal_altered"         "personal_loan_bal_altered"
## [17] "atm_withdrawls_cnt"        "atm_deposits_cnt"
## [19] "branch_visit_cnt"          "phone_banker_cnt"
## [21] "mobile_bank_cnt"           "online_bank_cnt"
## [23] "direct_mail_cnt"           "direct_email_cnt"
## [25] "direct_phone_cnt"
```

```
print('Mortgage Flag')
```

```
## [1] "Mortgage Flag"
```

```
summary(month_end_balances$mortgage_flag)
```

```
##   0   1
## 235  65
```

Example: Predicting whether someone will have a morgage

Looks like there are 65 people with morgages in the dataset and 235 without. What if you wanted to predict who had a mortgage and who didn't? There are a lot of different algorithms we could use. One of the easiest to use that yields good results is called a random forest. For our purposes at the moment it is enough to know that random forests is an ensemble based machine learning algorithm that can be used to predict an outcome we are interested in. Let's see how it can work.

```
library(randomForest)
```

```
## randomForest 4.6-10
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
fit <- randomForest(as.factor(mortgage_flag) ~ branch_visit_cnt + online_bank_cnt + direct_phone_cnt + a
                    data=month_end_balances,
                    importance=TRUE,
                    ntree=2000)
```

The above code creates our first model (fit). We had to specify the data and what we were prediction (mortgage_flag). We then had to give it what we want to use to predict mortgage flag after the ~. How do we see how we are doing? Welp. There is a convienent out of bag performance we can examine. For now we can use this metric as an estimate of performance.

```
print(fit)
```

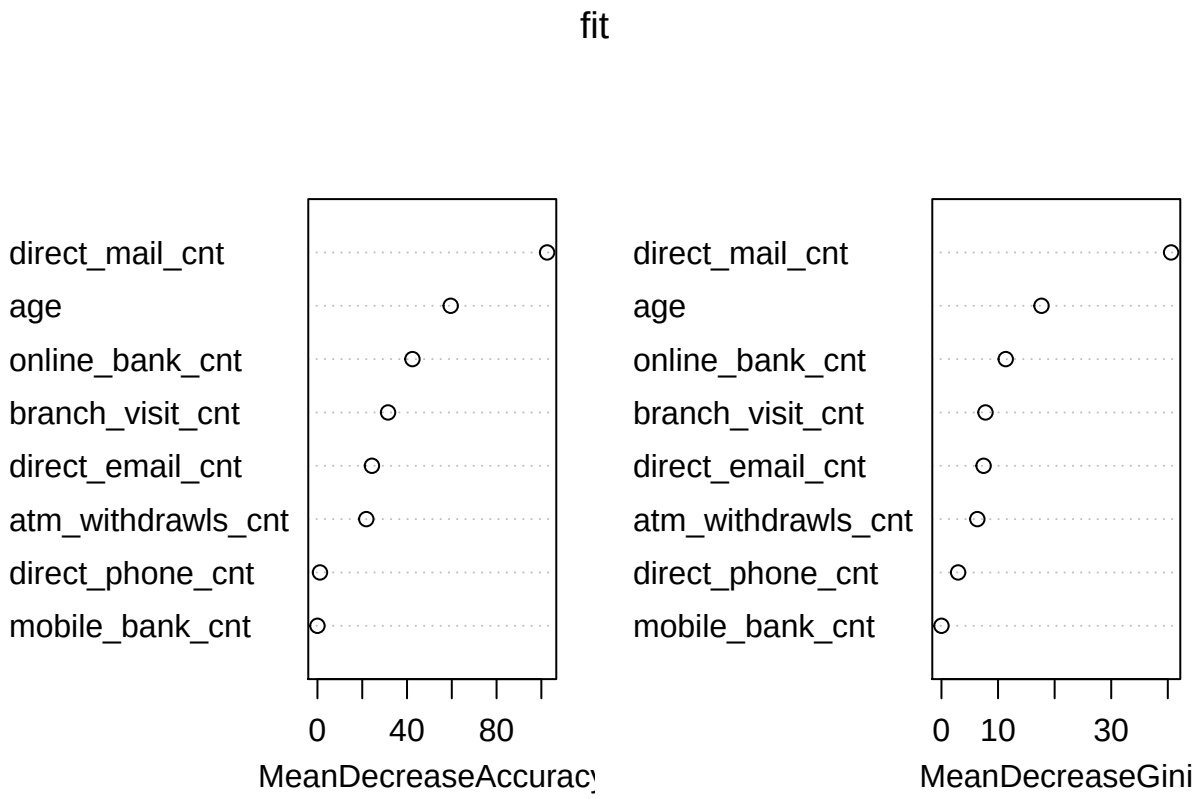
```
##
## Call:
##  randomForest(formula = as.factor(mortgage_flag) ~ branch_visit_cnt +      online_bank_cnt + direct_p
##                Type of random forest: classification
##                      Number of trees: 2000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 8%
## Confusion matrix:
##     0  1 class.error
## 0 232  3  0.01276596
## 1  21 44  0.32307692
```

It says we got an out of bag estimate of error of approximately 8%. This means we are wrong 8% of the time. The confusion matrix below shows where we went wrong broken up by class (0 = no mortgage or 1 =

mortgage). We always need to ask ourself if we are doing better than guessing. If we guessed that no one had a mortgage then we would get all 65 of those wrong, so we would have an error of $65/(65+235) = 22\%$. Awesome! With a couple of lines we have a decent model. You can play around with the parameters.

Now everyone loves a graph, so a cool thing about random forest is you can see how important a variable is to prediction:

```
varImpPlot(fit)
```



This shows us that things like direct phone cnt and mobile bank cnt aren't that important for this prediction.

What about using this kind of prediction for your idea?

Well. What you need to do is figure out exactly what you are trying to predict and how that would be useful for either the consumer or for wells fargo. But in terms of nuts and bolts. All you need is all the data you want in a single data frame and then you'll need to pick the column you want to predict. Think recommendation systems.

Here is my concrete suggestion for everyone in week 2 of the competition. Pick something from the data that is related to your idea and try to predict it :) You will probably have to do some munging of the data to get exactly what you want, but I'll be here on Facebook to help you out.