Machine Learning 2018
Introduction to Course Projects

Kien C Nguyen

November 21, 2018

VEF
ACADEMY

1. Heritage Health Prize
2. Santander Product Recommendation
3. Home Credit Default Risk

Heritage Health Prize

- More than 71 million people in the United States are admitted to hospitals each year, according to the latest survey from the American Hospital Association.
- Studies have concluded that in 2006 well over $30 billion was spent on unnecessary hospital admissions.
- The Heritage Provider Network (HPN) believes that we can identify earlier those most at risk and ensure they get the treatment they need.

Source: `https://www.kaggle.com/c/hhp`

Heritage Health Prize

- In this project, we will build an algorithm that predicts how many days a patient will spend in a hospital in the next year.
- Using this information, health care providers can develop new care plans and strategies to reach patients before emergencies occur, thereby reducing the number of unnecessary hospitalizations.
- This will result in increasing the health of patients while decreasing the cost of care.

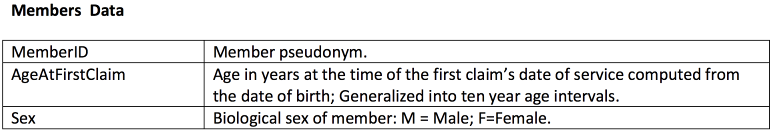
Source: `https://www.kaggle.com/c/hhp`

The Data Set

- There is missing data because this is real data that was pulled out of production systems
- There are no names or identities of any of the patients – therefore risk assessments had to be done with estimates and simulations
- The competition data set represents a small sample of HPN members – the sub-sampling has a big impact on re-identification risk

Source:
`https://foreverdata.org/1015/deidentification.pdf`

Member data

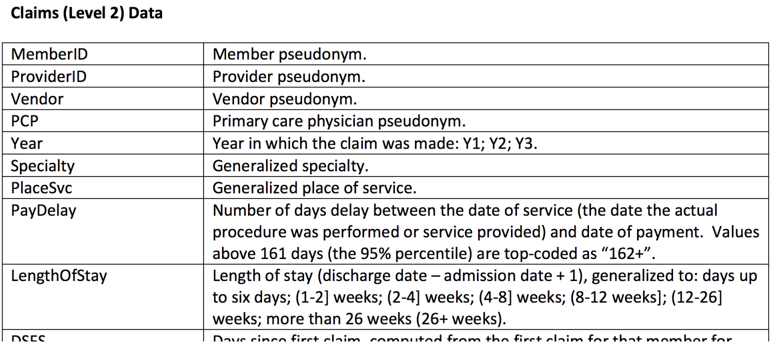


Members Data

MemberID	Member pseudonym.
AgeAtFirstClaim	Age in years at the time of the first claim's date of service computed from the date of birth; Generalized into ten year age intervals.
Sex	Biological sex of member: M = Male; F=Female.

Source: `https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf`

Claim data

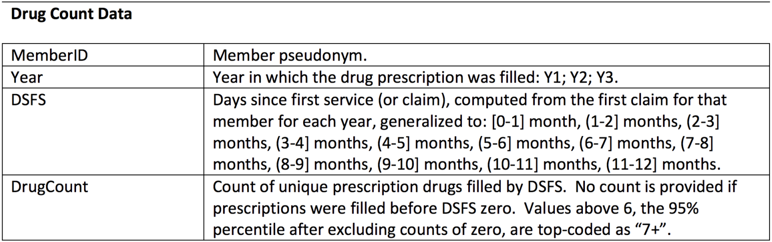


Claims (Level 2) Data

MemberID	Member pseudonym.
ProviderID	Provider pseudonym.
Vendor	Vendor pseudonym.
PCP	Primary care physician pseudonym.
Year	Year in which the claim was made: Y1; Y2; Y3.
Specialty	Generalized specialty.
PlaceSvc	Generalized place of service.
PayDelay	Number of days delay between the date of service (the date the actual procedure was performed or service provided) and date of payment. Values above 161 days (the 95% percentile) are top-coded as “162+”.
LengthOfStay	Length of stay (discharge date – admission date + 1), generalized to: days up to six days; (1-2] weeks; (2-4] weeks; (4-8] weeks; (8-12 weeks]; (12-26] weeks; more than 26 weeks (26+ weeks).
DSFS	Days since first claim, computed from the first claim for that member for

Source: `https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf`

Drug count data

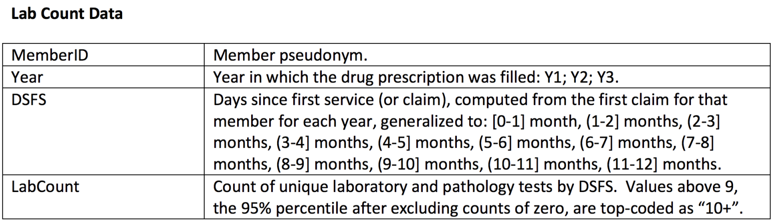


Drug Count Data

MemberID	Member pseudonym.
Year	Year in which the drug prescription was filled: Y1; Y2; Y3.
DSFS	Days since first service (or claim), computed from the first claim for that member for each year, generalized to: [0-1] month, (1-2] months, (2-3] months, (3-4] months, (4-5] months, (5-6] months, (6-7] months, (7-8] months, (8-9] months, (9-10] months, (10-11] months, (11-12] months.
DrugCount	Count of unique prescription drugs filled by DSFS. No count is provided if prescriptions were filled before DSFS zero. Values above 6, the 95% percentile after excluding counts of zero, are top-coded as "7+".

Source: https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf

Lab count data

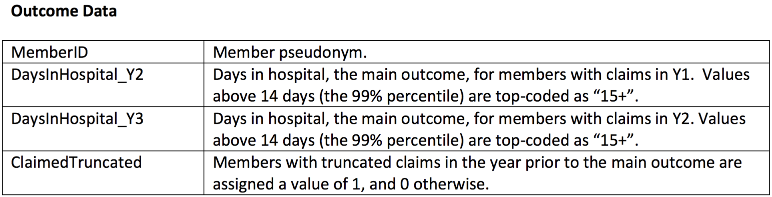


Lab Count Data

MemberID	Member pseudonym.
Year	Year in which the drug prescription was filled: Y1; Y2; Y3.
DSFS	Days since first service (or claim), computed from the first claim for that member for each year, generalized to: [0-1] month, (1-2] months, (2-3] months, (3-4] months, (4-5] months, (5-6] months, (6-7] months, (7-8] months, (8-9] months, (9-10] months, (10-11] months, (11-12] months.
LabCount	Count of unique laboratory and pathology tests by DSFS. Values above 9, the 95% percentile after excluding counts of zero, are top-coded as "10+".

Source: https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf

Outcome data



Outcome Data

MemberID	Member pseudonym.
DaysInHospital_Y2	Days in hospital, the main outcome, for members with claims in Y1. Values above 14 days (the 99% percentile) are top-coded as "15+".
DaysInHospital_Y3	Days in hospital, the main outcome, for members with claims in Y2. Values above 14 days (the 99% percentile) are top-coded as "15+".
ClaimedTruncated	Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise.

Source: https://foreverdata.org/1015/content/Data_Dictionary_release3.pdf

- In this competition, you are provided with 1.5 years of customers behavior data from Santander bank to predict what new products customers will purchase.
- The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", etc.
- You will predict what additional products a customer will get in the last month, 2016-06-28, in addition to what they already have at 2016-05-28.
- These products are the columns named: $ind\_(xyz)\_ult1$, which are the columns #25 - #48 in the training data.

Source: `https://www.kaggle.com/c/santander-product-recommendation`

Home Credit Default Risk

- Many people struggle to get loans due to insufficient or non-existent credit histories.
- And, unfortunately, this population is often taken advantage of by untrustworthy lenders.
- Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.
- In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities.

Source:
`https://www.kaggle.com/c/home-credit-default-risk`

Home Credit Default Risk

- While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data.
- Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Source:
`https://www.kaggle.com/c/home-credit-default-risk`

application_{train | test }.csv
This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET). One row represents one loan in our data sample.
bureau.csv
All client's previous credits provided by other financial institutions that were reported to Credit Bureau For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
bureau_balance.csv
Monthly balances of previous credits in Credit Bureau. This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
Source:
`https://www.kaggle.com/c/home-credit-default-risk`

POS_CASH_balance.csv

Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit. This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.

credit_card_balance.csv

Monthly balance snapshots of previous credit cards that the applicant has with Home Credit. This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (# loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.

previous_application.csv
All previous applications for Home Credit loans of clients who have loans in our sample. There is one row for each previous application related to loans in our data sample.
installments_payments.csv
Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample. There is a) one row for every payment that was made plus b) one row each for missed payment. One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
HomeCredit_columns_description.csv
This file contains descriptions for the columns in the various data files.

Column Descriptions

https://www.kaggle.com/c/home-credit-default-risk/data

References

[1] Heritage Health Prize – https://www.kaggle.com/c/hhp

[2] Heritage Health Prize Contest Data – https://foreverdata.org/1015/?fbclid=IwAR12c6zFFDOMc3Gm9001CXmTQZavgJ03tJc8_wjMQ_vI4pURiGumixZ3FQE

[3] Santander Product Recommendation – https://www.kaggle.com/c/santander-product-recommendation

[4] Home Credit Default Risk – https://www.kaggle.com/c/home-credit-default-risk