MINIMIZE RISK USING MACHINE LEARNING METHODS

Instructor: Prof. M. Daneshmand

Submitted by: Jhao-Han Chen

TOPICS

- Business understanding
- Data understanding
- Data cleaning
- Modeling
- Evaluation
- Conclusion
- Reference

BUSINESS UNDERSTANDING - ORGANIZATION

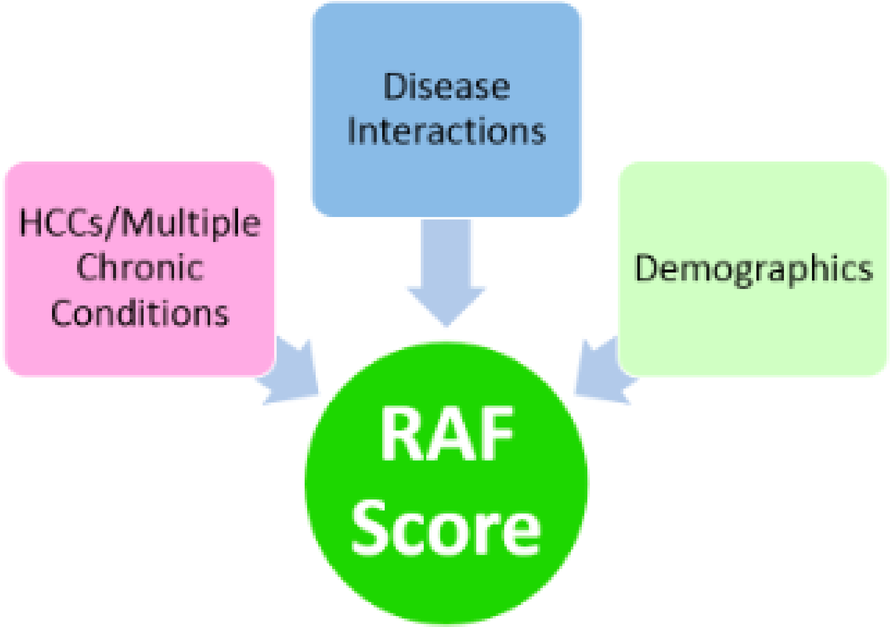
- The **Centers for Medicare & Medicaid Services** (**CMS**)

is a federal agency within the United States Department of Health and Human Services (HHS) that administers the Medicare program and works in partnership with state governments to administer Medicaid, the Children's Health Insurance Program (CHIP), and health insurance portability standards.

BUSINESS UNDERSTANDING – RAF SCORE

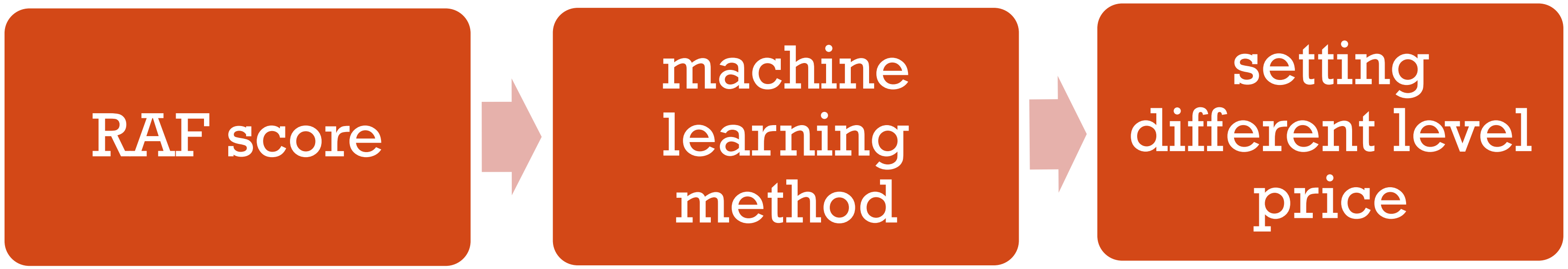
CMS initiated Hierarchical condition category(HCC) model which is a risk-adjustment model designed to estimate future health care costs for patient in 2004.

Hierarchical condition category coding helps communicate patient complexity and paint a picture of the whole patient. In addition to helping predict health care resource utilization, RAF scores are used to risk adjust quality and cost metrics. By accounting for difference in patient complexity, quality and cost performance can be more appropriately measured.



BUSINESS UNDERSTANDING – PROBLEM

- People nowadays buy insurances to prevent risk from happening, especially for medical insurance.
- In order to save cost, insurance companies have to evaluate patients' condition before setting a price.
- Assigning patients with different levels price is a big issue for insurance company.
- Project goal:



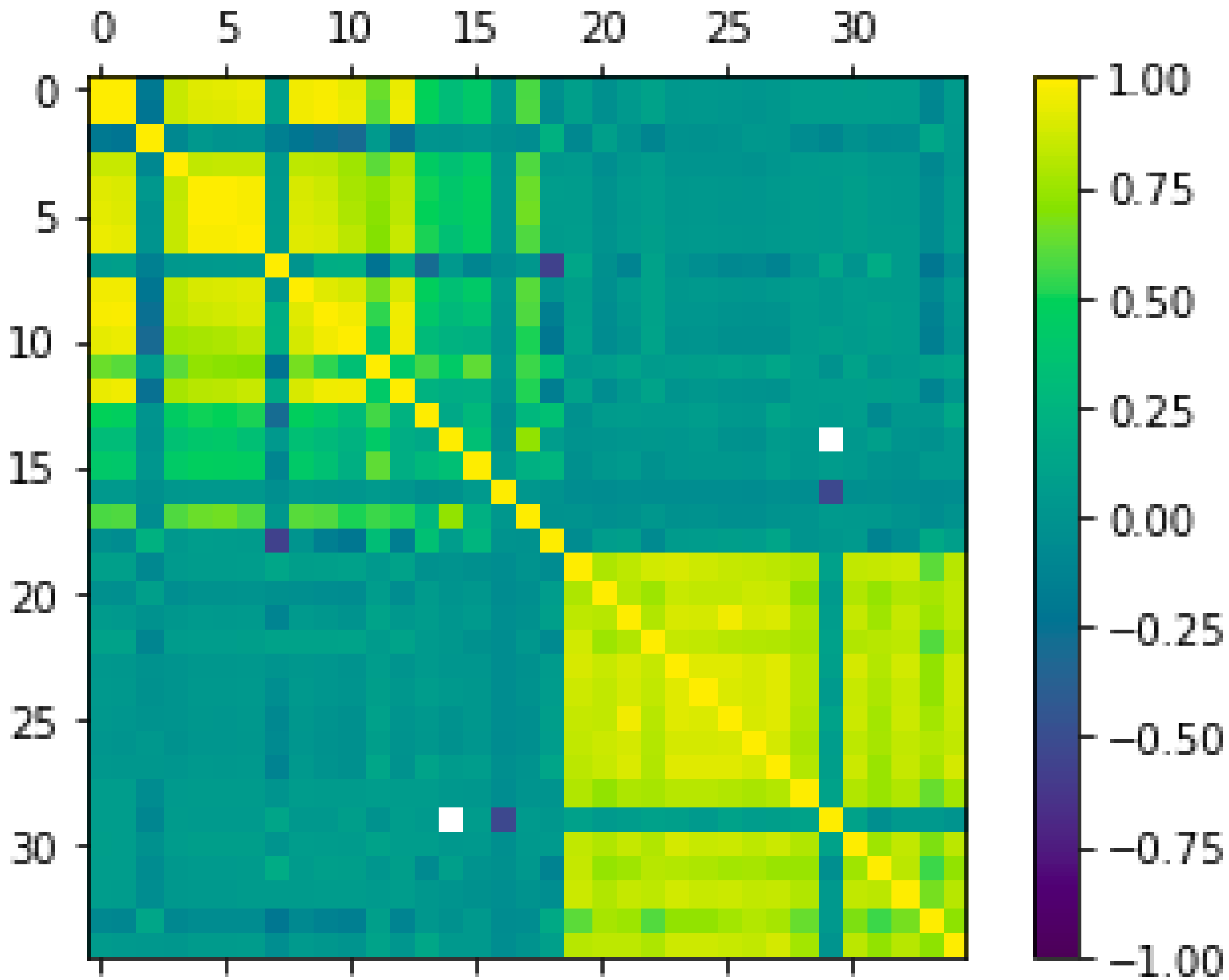
DATA UNDERSTANDING- LOAD DATA

- My dataset is download from Kaggle. The data source is come from CMS official website
- Using python to load data and to conduct end to end analyze
- Columns contain 28 float type, 9 integer type, and 4 object type

	Provider ID	Facility Name	Street Address	City	State	Zip Code	Total Stays	Distinct Beneficiaries Per Provider	Average Length of Stay (Days)	Total SNF Charge Amount	...	Percent of Beneficiaries with COPD	Percent of Beneficiaries with Depression	Percent of Beneficiaries with Diabetes	Perce Benef with Hype
0	10022	CHEROKEE MEDICAL CENTER	400 NORTHWOOD DR	CENTRE	AL	35960	95	85	13.9	3787309	...	39.0	52.0	45.0	
1	10032	WEDOWEE HOSPITAL	209 NORTH MAIN STREET	WEDOWEE	AL	36278	20	19	10.6	436623	...	NaN	21.0	53.0	
2	10044	MARION REGIONAL MEDICAL CENTER	1256 MILITARY STREET SOUTH	HAMILTON	AL	35570	164	144	15.4	5906115	...	44.0	42.0	50.0	
3	10045	FAYETTE MEDICAL CENTER	1653 TEMPLE AVENUE NORTH	FAYETTE	AL	35555	124	110	16.0	2748027	...	42.0	34.0	38.0	
4	10058	BIBB MEDICAL CENTER	208 PIERSON AVE	CENTREVILLE	AL	35042	90	85	17.4	1679414	...	34.0	40.0	56.0	

5 rows × 41 columns

```
Provider ID                                          15026 non-null int64
Facility Name                                        15026 non-null object
Street Address                                       15026 non-null object
City                                                 15026 non-null object
State                                                15026 non-null object
Zip Code                                             15026 non-null int64
Total Stays                                          15026 non-null int64
Distinct Beneficiaries Per Provider                  15026 non-null int64
Average Length of Stay (Days)                        15026 non-null float64
Total SNF Charge Amount                              15026 non-null int64
Total SNF Medicare Allowed Amount                    15026 non-null int64
Total SNF Medicare Payment Amount                    15026 non-null int64
Total SNF Medicare Standard Payment Amount           15026 non-null int64
Average Age                                          15026 non-null int64
Male Beneficiaries                                   13454 non-null float64
Female Beneficiaries                                 13454 non-null float64
Nondual Beneficiaries                                12205 non-null float64
Dual Beneficiaries                                   12205 non-null float64
White Beneficiaries                                  14636 non-null float64
Black Beneficiaries                                  8853 non-null float64
Asian Pacific Islander Beneficiaries                 9521 non-null float64
Hispanic Beneficiaries                               7889 non-null float64
American Indian or Alaska Native Beneficiaries       11431 non-null float64
Other/ Unknown Beneficiaries                         5263 non-null float64
Average HCC Score                                    15026 non-null float64
Percent of Beneficiaries with Atrial Fibrillation    15026 non-null float64
Percent of Beneficiaries with Alzheimer's            14088 non-null float64
Percent of Beneficiaries with Asthma                 15026 non-null float64
Percent of Beneficiaries with Cancer                 15026 non-null float64
Percent of Beneficiaries with CHF                    14816 non-null float64
Percent of Beneficiaries with Chronic Kidney Disease 14504 non-null float64
Percent of Beneficiaries with COPD                   14999 non-null float64
Percent of Beneficiaries with Depression             14553 non-null float64
Percent of Beneficiaries with Diabetes               14884 non-null float64
Percent of Beneficiaries with Hyperlipidemia         13024 non-null float64
Percent of Beneficiaries with Hypertension           218 non-null float64
Percent of Beneficiaries with IHD                    13950 non-null float64
Percent of Beneficiaries with Osteoporosis           15024 non-null float64
Percent of Beneficiaries with RA/OA                  14193 non-null float64
Percent of Beneficiaries with Schizophrenia          14879 non-null float64
Percent of Beneficiaries with Stroke                 15026 non-null float64
```



DATA UNDERSTANDING- CORRELATION HEATMAP

- The lighter the column the higher correlated the columns
- Columns of Percent of Beneficiaries with illnesses and columns of Beneficiaries for Underprivileged Groups are highly correlated, therefore I will consider drop them while cleaning the data

DATA CLEANING – DROP COLUMNS

- Drop object column(“Facility Name “,” Street Address”, “City “, “State “)
- Drop not important column('Provider ID', 'Zip Code’)
- Drop “Total Stays” retain “Average Length of Stay (Days)”
- Drop 'American Indian or Alaska Native Beneficiaries' because 75% of the column are 0
- drop 'Average HCC Score' which is used for target cell

DATA CLEANING – MISSING VALUE

- Drop columns which have too many missing value(“Black Beneficiaries “, “Hispanic Beneficiaries “, “Other/ Unknown Beneficiaries “, “Percent of Beneficiaries with Hypertension “)

```
#drop columns which have too many missing value
data=data[data.columns[data.isnull().mean()<0.4]]
```

- Convert rest missing values to 0(consider null value as no beneficiaries)

```
#fill NuLL value to 0
data=data.fillna(0)
```

DATA CLEANING - MULTICOLLINEARITY

- Drop highly correlated column(17 columns)
- Drop highly correlated columns of Beneficiaries for underprivileged groups('Asian Pacific Islander Beneficiaries')
- Drop highly correlated columns of Percent of Beneficiaries with illnesses (5 columns)

```
# Create correlation matrix
corr_matrix = data.corr().abs()

# Select upper triangle of correlation matrix
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool))

# Find index of feature columns with correlation greater than 0.87
to_drop = [column for column in upper.columns if any(upper[column] > 0.87)]

#drop highly correlated column
data.drop(to_drop,inplace=True, axis=1)
```

DATA CLEANING - OUTLIERS

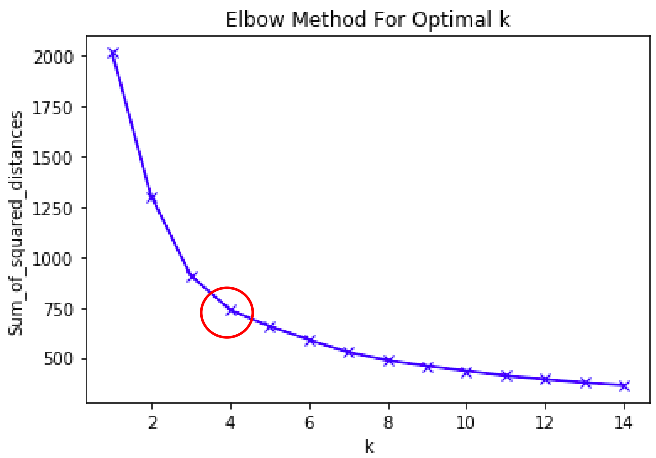
- delete 'Dual Beneficiaries', 'Percent of Beneficiaries with Osteoporosis' outliers(544 rows)

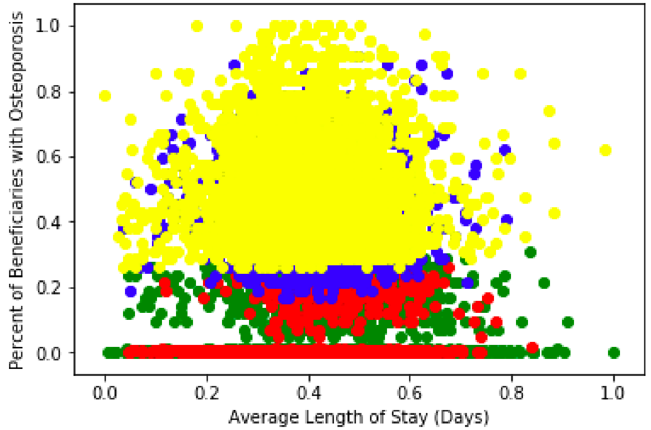
	Average Length of Stay (Days)	Total SNF Charge Amount	Average Age	Dual Beneficiaries	Average HCC Score	Percent of Beneficiaries with Osteoporosis
q1	23.2	1027862.75	76.0	15.0	2.050	0.167670
q3	31.8	3820889.25	82.0	64.0	2.760	17.000000
iqr	8.6	2793026.50	6.0	49.0	0.710	16.832330
upper_limit	44.7	8010429.00	91.0	137.5	3.825	42.248495
lower_limit	10.3	-3161677.00	67.0	-58.5	0.985	-25.080825

MODELING - DATA

- Cleaned data contain 14482 rows , 5 columns
- Normalized data

	Average Length of Stay (Days)	Total SNF Charge Amount	Average Age	Dual Beneficiaries	Percent of Beneficiaries with Osteoporosis
0	0.188768	0.038278	0.688889	0.109489	0.333333
1	0.137285	0.004140	0.777778	0.000000	0.380952
2	0.212168	0.059865	0.666667	0.306569	0.380952
3	0.221529	0.027689	0.711111	0.270073	0.309524
4	0.243370	0.016802	0.644444	0.226277	0.261905

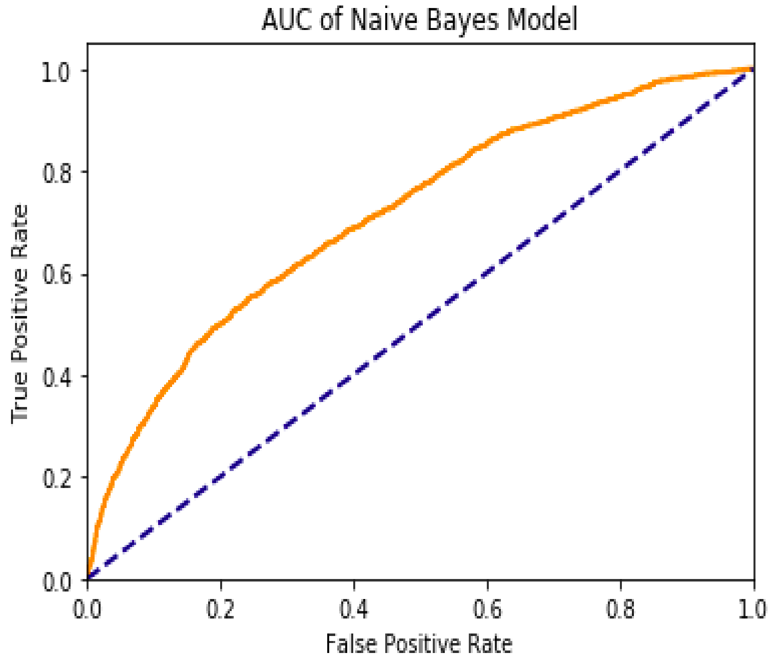


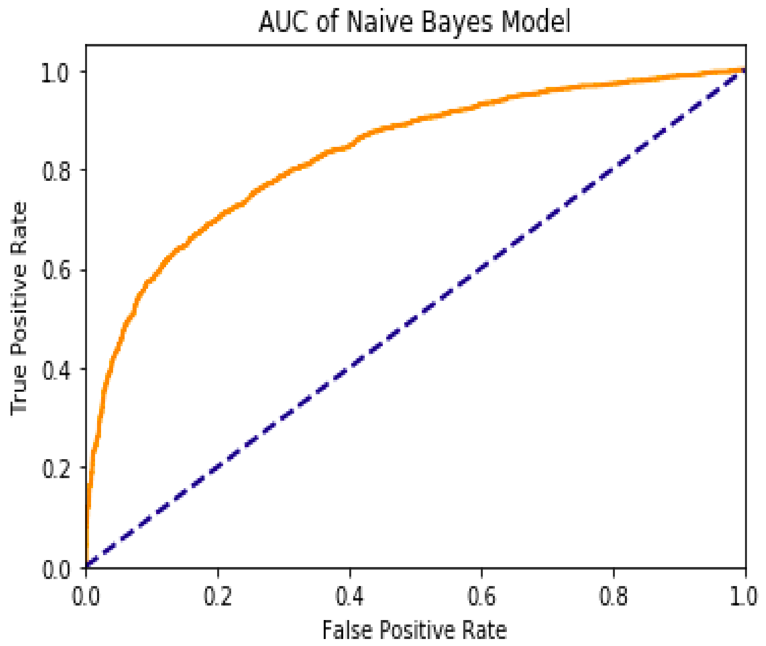


MODELING - CLUSTERING

- Find optimal k=4
- The visualized result seem to have two clusters
- For the target label, I divided "Average HCC Score" into four parts by 1,2,3 quantile point, which I can use it on evaluating the clustering model.
- Cluster centroid:

cluster	Average Length of Stay (Days)	Total SNF Charge Amount	Average Age	Dual Beneficiaries	Percent of Beneficiaries with Osteoporosis
0	0.410735	0.016185	0.704446	0.123572	0.028952
1	0.406674	0.042638	0.674025	0.506882	0.017442
2	0.388124	0.047780	0.686249	0.577170	0.424803
3	0.391260	0.019432	0.743509	0.156625	0.512155





MODELING – NAÏVE BAYES & SVM

- First, I split scaled data into 0.7 training and 0.3 testing data
- Second, I divided "Average HCC Score" into two parts to be the target cell which I can use on evaluating the model.
- Last, I fit the model with five fold cross validation and plot ROC curve

EVALUATION – F1 SCORE & AUC

Clustering

	precision	recall	f1-score	support
danger_level	0.38	0.34	0.36	3695
high_risk	0.33	0.21	0.26	3609
low_risk	0.40	0.37	0.38	3572
normal_risk	0.25	0.39	0.31	3606
micro avg	0.33	0.33	0.33	14482
macro avg	0.34	0.33	0.33	14482
weighted avg	0.34	0.33	0.33	14482

SVM

Test data set average f1 score:
0.7554698840987291

Test data set average auc:
0.8334285248754447

Naïve Bayes

Test data set average f1 score:
0.6197032259584737

Test data set average auc:
0.7279020608784498

EVALUATION

- Clustering seem not to be a good method for this dataset. We can tell after seeing the cluster visualization.
- SVM is the best model in classifying this dataset and predict the assigned price level, which achieve auc of 0.8334

model	f1 score	auc
Clustering	0.33	
Naïve Bayes	0.6197	0.7279
SVM	0.7555	0.8334

CONCLUSION

- Using SVM classified the medical data set successfully which achieve auc 0.8334
- Insurance companies can refer to their customers' medical information and use SVM model to predict the risk of selling insurance, and minimize the cost lastly.
- Clustering is not a good method while dealing with medical data. Perhaps there are large differences between each medical facilities lead to low accuracy of predicted result. More researches should be done on dealing with medical data.

REFERENCE

- https://en.wikipedia.org/wiki/Centers_for_Medicare_and_Medicaid_Services
- https://www.aafp.org/practice-management/payment/coding/hcc.html
- https://www.medirevv.com/blog/what-is-hcc-coding-understanding-todays-risk-adjustment-model
- https://www.kaggle.com/cms/medicare-skilled-nursing-facility-provider-reports#medicare-skilled-nursing-facility-snf-provider-by-rug-aggregate-report-cy-2014.csv
- https://www.cms.gov/about-cms/about-cms.html
- https://www.aafp.org/fpm/2016/0900/p24.html