The Housing Market Health Evaluation based on Zillow Research Data

Team 14: Bowen Lu, Hao Cheng

I. Introduction

Zillow's Market Health Index uses up to 10 market indicators to measure the health of markets relative to similar markets across the country on a scale of 1 to 10 down to the ZIP code, neighborhood, city, county, metro and state level. The feature is a "key component" of Zillow's new local information pages, which aim to offer users more data, Zillow said.

If a market's Market Health Index score is 8, that means Zillow considers it to be healthier than 80 percent of all comparable areas that it covers. Zillow said that a market that scores low on its health meter may not necessarily be "performing poorly" since a market's score is assigned based on how it stacks up against other similar markets.

"A low Market Health Index score does not necessarily indicate that a market is performing poorly, only that other markets are experiencing factors like higher home value appreciation or lower foreclosure activity," Zillow said.

Zillow scores the health of a market only if it has data on at least five of all the indicators that it may use for the index. For instance, some of the indicators are:

- month-over-month change in the Zillow Home Value Index (ZHVI).
- year-over-year change in ZHVI.
- percent change in one-year ZHVI forecast.
- percentage of homes selling for a gain.
- the number of days listings spend on Zillow, adjusted for seasonality and for deviations from historic norms.
- the percentage of mortgage holders with negative equity.

- the percentage of mortgage holders' delinquent on their loans.

II. Data Preparation

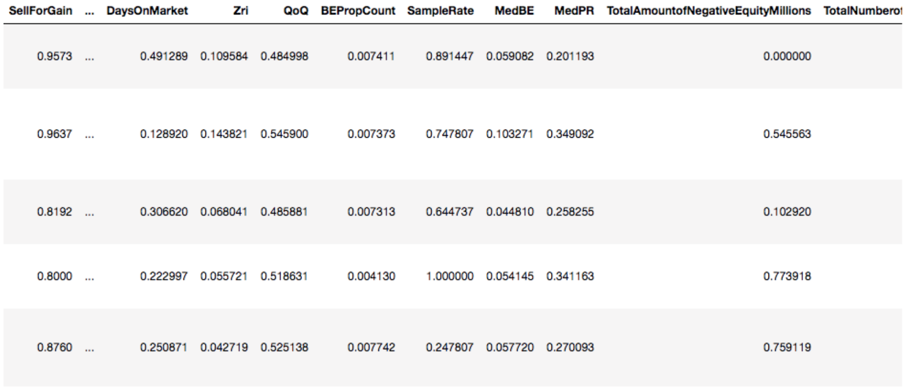
1. Data Description

We obtained our dataset from *https://www.zillow.com/research/data/* . The dependent variable is the Market Health Index and the other 17 variables are the independent variables.

2. Data Preparation

Then, we attempted to clean the raw data. Firstly, we use R program dealing with the missing values. If the missing values take up more than 50%, we dropped those variables' columns. Otherwise, we substituted the missing values by the mode of K-Nearest-Numbers. After filling the missing values, we kept 17 independent variables at last.

Secondly, we had to standardize the variables to better compare the relationship between themselves and also to make them much more normal. (Figure 2)



SellForGain	...	DaysOnMarket	Zri	QoQ	BEPropCount	SampleRate	MedBE	MedPR	TotalAmountofNegativeEquityMillions	TotalNumberot
0.9573	...	0.491289	0.109584	0.484998	0.007411	0.891447	0.059082	0.201193	0.000000	
0.9637	...	0.128920	0.143821	0.545900	0.007373	0.747807	0.103271	0.349092	0.545563	
0.8192	...	0.306620	0.068041	0.485881	0.007313	0.644737	0.044810	0.258255	0.102920	
0.8000	...	0.222997	0.055721	0.518631	0.004130	1.000000	0.054145	0.341163	0.773918	
0.8760	...	0.250871	0.042719	0.525138	0.007742	0.247807	0.057720	0.270093	0.759119	

Figure 2

Next, we have generated the P-P Plot to assess the normality. Fortunately, all the variables are distributed normally after standardizing. So, we didn't need to implement transformation at all. Some results are attached below.

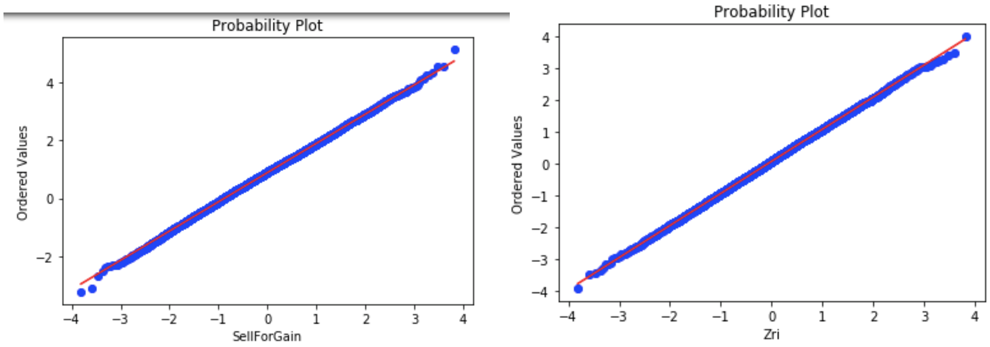


Figure 3

III. Multiple Regression

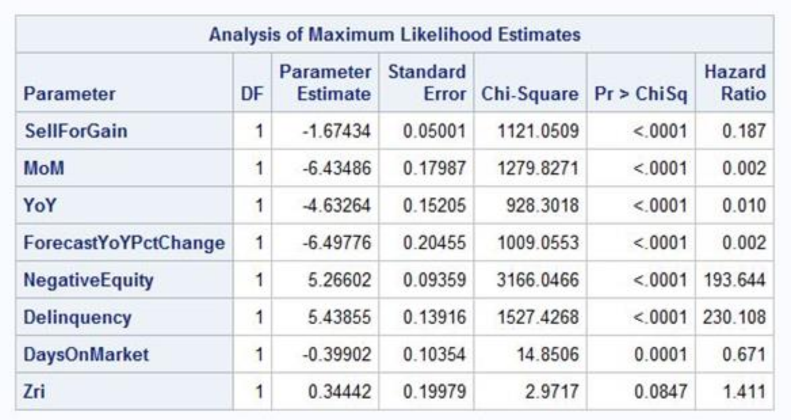
1. Methodology

In order to evaluate Market Health Index, we built a multiple regression model to reflect the quantitative relation between Market Health Index and the seventeen independent variables mentioned before. The model is linear but multidimensional.

Besides, a stepwise variable selection is conducted to eliminate the variables that not have significant impact on Market Health Index. The F-score of each independent variable and its corresponding p-value are calculated. For p-value, we set the cutoff point for entry (0.15) and the cutoff point for remove (0.30). In each step, the unselected variable with $p<0.15$ and highest F-score will be selected; meanwhile, the selected variables with $p>0.30$ will be removed. This iteration end up with no variable to be entered or removed.

2. Result

Table 3.1 is the result of the stepwise variable selection. Based on the maximum likelihood estimates, eight variables are selected. The following regression and classification mainly use them for the analysis.

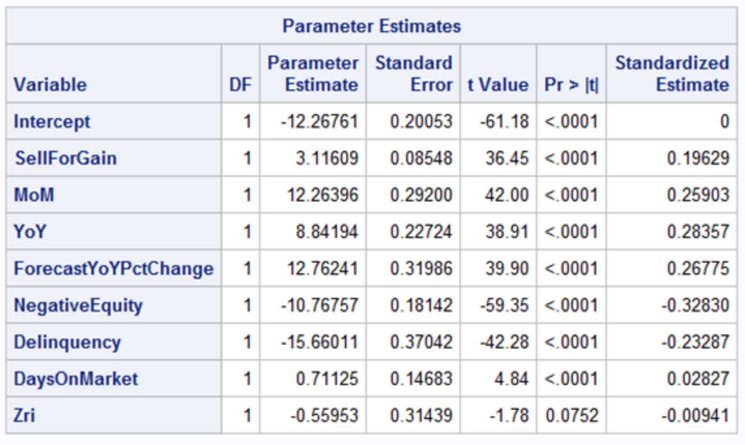


Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
SellForGain	1	-1.67434	0.05001	1121.0509	<.0001	0.187
MoM	1	-6.43486	0.17987	1279.8271	<.0001	0.002
YoY	1	-4.63264	0.15205	928.3018	<.0001	0.010
ForecastYoYPctChange	1	-6.49776	0.20455	1009.0553	<.0001	0.002
NegativeEquity	1	5.26602	0.09359	3166.0466	<.0001	193.644
Delinquency	1	5.43855	0.13916	1527.4268	<.0001	230.108
DaysOnMarket	1	-0.39902	0.10354	14.8506	0.0001	0.671
Zri	1	0.34442	0.19979	2.9717	0.0847	1.411

Table 3.1

Table 3.2 display the parameter estimates of the eight selected variables and the intercept in the multiple regression model. Their p-value are smaller than 0.05, so the null hypothesis is rejected, which means the eight variables all have significant impact on Market Health Index. The standardized estimate refers to the correlation coefficient between the predictor variable and Market Health Index. The regression equation is obtained according to the parameter estimates:

$$\begin{aligned}\text{MarketHealthIndex} = {} & -12.27 + 3.12 * \text{SellForGain} + 12.26 * \text{MoM} + 8.84 * \text{YoY} + 12.76 * \\ & \text{ForecastYoYPctChange} - 10.77 * \text{NegativeEquity} - 15.66 * \\ & \text{Delinquency} + 0.71 * \text{DaysOnMarket} - 0.56 * \text{Zri}\end{aligned}$$

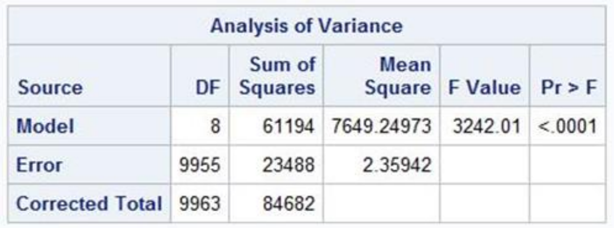


Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-12.26761	0.20053	-61.18	<.0001	0
SellForGain	1	3.11609	0.08548	36.45	<.0001	0.19629
MoM	1	12.26396	0.29200	42.00	<.0001	0.25903
YoY	1	8.84194	0.22724	38.91	<.0001	0.28357
ForecastYoYPctChange	1	12.76241	0.31986	39.90	<.0001	0.26775
NegativeEquity	1	-10.76757	0.18142	-59.35	<.0001	-0.32830
Delinquency	1	-15.66011	0.37042	-42.28	<.0001	-0.23287
DaysOnMarket	1	0.71125	0.14683	4.84	<.0001	0.02827
Zri	1	-0.55953	0.31439	-1.78	0.0752	-0.00941

Table 3.2

3. Analysis and model performance

The MANOVA is applied to measure the performance of the regression model by analyzing both the variance of regression model and the variance of residual (error). In table 3.3, the degree of freedom of model is 8 because of eight independent variables, and the degree of freedom of error is N−P−1=9955. MSR equals to SSR divided by d.f. of model, while MSE equals to SSE divided by d.f. of error. As the ratio of MSR and MSE, F-value have less than 0.001 probability that Pr>$F_{0.05}$. This confirms that the linear combination of eight independent variables has significant impact on Market Health Index.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	61194	7649.24973	3242.01	<.0001
Error	9955	23488	2.35942		
Corrected Total	9963	84682			

Table 3.3

RMSE and R^2 (adjusted or not) are calculated to measure the performance of the regression model. RMSE is the square root of MSE, reflecting the impact of residuals on Y. R^2 is also called *coefficient of determination*, indicating the strength of association between Y and the set of X. For instance, $100R^2\%$ is the percent of variance of Y explained by the set of X. Table 3.4 compares RMSE and R^2 and adjusted R^2 of the regression model with all variables and the one with eight selected variables.

	Model of 17 variables	Model of 8 variables
RMSE	1.53654	1.53604
R^2	0.7227	0.7226
Adj R^2	0.7222	0.7224

Table 3.4

After the variable selection, RMSE decreases a little, so the impact of residuals on Y gets weaker. R^2 reduces because as long as the number of independent variables decreases, R^2 certainly decreases. However, the adjusted R^2 increases, which shows the improvement of the regression model performance.

IV. Classification

1. Methodology

Wikipedia says classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. In this project, therefore, a binary variable MarketHealth is defined as the targeted variable for classification, according to the Market Health Index: If MHI ≥ 6, MarketHealth = 1 (means a healthy market); If MHI < 6, MarketHealth = 0 (means an unhealthy market).

As for the classification models, this project used three of them:

1) Linear Discriminant Analysis

   Build a linear combination of features that characterizes or separates two classes of records. In practical, for a new record, the value of one variable multiplies the discriminant function coefficient of this variable, then sum up the product of all variables to calculate the score of each class. The new record will be classified to the class with higher score.

   Please notice that the frequency of MarketHealth = 1 and MarketHealth = 0 is similar (4484 vs. 5480), so the prior probability of each classes is regarded as 0.5. Otherwise a product with $\log(P_i)$ should be added to the classification function.

2) Logistic Regression

   Logistic Regression starts with the posterior probability from discriminant analysis.

$$P_z = \frac{1}{1 + e^{C-Z}}$$

   The odd of the event that a new record belongs to class Z_i is an exponent function. As a result, the natural logarithm of odds can be regarded as a linear classifier.

$$ln\,(Odds) = ln\left(\frac{P_z}{(1-P_z)}\right) = \boldsymbol{\beta X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = Z - C$$

The coefficients of functions are also estimated by maximum likelihood method and an interactive process called Iterative Weighted Least Squares.

3) Naïve Bayes

Naïve Bayes a family of algorithms based on the Bayes Theorem:

$$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

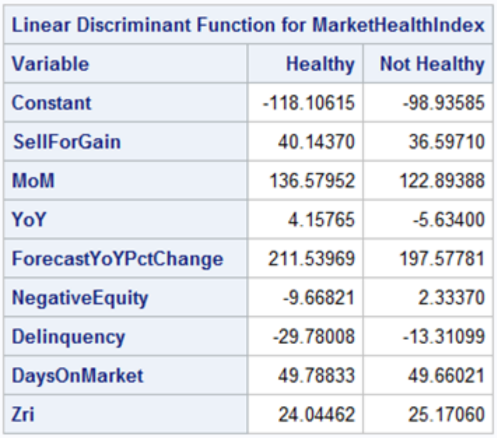
All Naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Therefore, it infers that:

$$p(C|F_1,\dots,F_n) = \frac{p(C)\prod_1^n(F_i|C)}{p(F_1,\dots,F_n)}$$

The class that a new record belongs to is determined by the largest posterior probability of class C_i given the new record's features (F_1, … , F_n).

2. Result

Table 4.1 shows the independent variables' coefficient of the discriminant function of the dependent variable. Supposing that *Hoboken* is a new record to be classified, with the value of eight features: (0.8889, 0.39601617, 0.551712509, 0.557066754, 0.210198521, 0.00130719, 0.254355401, 0.069465082).



Linear Discriminant Function for MarketHealthIndex		
Variable	Healthy	Not Healthy
Constant	-118.10615	-98.93585
SellForGain	40.14370	36.59710
MoM	136.57952	122.89388
YoY	4.15765	-5.63400
ForecastYoYPctChange	211.53969	197.57781
NegativeEquity	-9.66821	2.33370
Delinquency	-29.78008	-13.31099
DaysOnMarket	49.78833	49.66021
Zri	24.04462	25.17060

Table 4.1

Import the value vector to calculate the score of different class. The score of “Healthy” (MarketHealth = 1) is 104.0719, with the posterior probability of classifying as “Healthy” being 0.5022. In contrast, the score of “Unhealthy” (MarketHealth = 0) is 104.0638, with the posterior probability of classifying as “Unhealthy” being 0.4977. As a result, the record *Hoboken* is classified as a health market of housing.

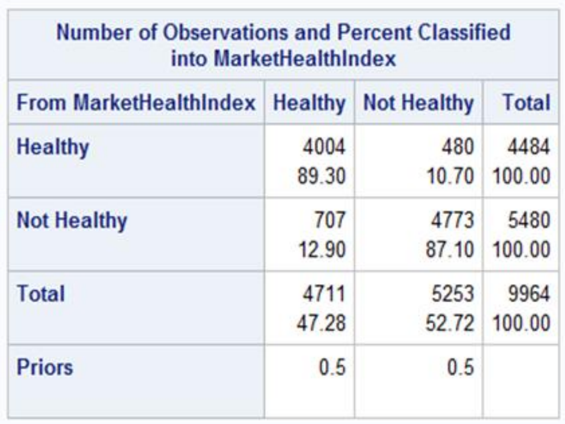
Due to the lack of space, the classifications of a single record operated by the logistic regression model and the Naïve Bayes model are no longer to be demonstrated in detail. The classification accuracy and other performance index of the entire dataset are more important, which will be analyzed in following part.

3. Analysis and model performance

The performance of classification models including the accuracy, confusion matrix and related evaluation index (precision, recall, f-score), and also ROC curve. They have some similarities so we choose a part of them to display for each model.

1) Linear Discriminant Analysis

Table 4.2 is the confusion matrix of the LDA model. The frequency of “True-Positive” is 4004; the frequency of “False-Positive” is 480; the frequency of “False-Negative” is 707; the frequency of “True-Negative” is 4773.



Number of Observations and Percent Classified into MarketHealthIndex

From MarketHealthIndex	Healthy	Not Healthy	Total
Healthy	4004 89.30	480 10.70	4484 100.00
Not Healthy	707 12.90	4773 87.10	5480 100.00
Total	4711 47.28	5253 52.72	9964 100.00
Priors	0.5	0.5	

Table 4.2

According to the confusion matrix, we can get that:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} = 0.8809$$

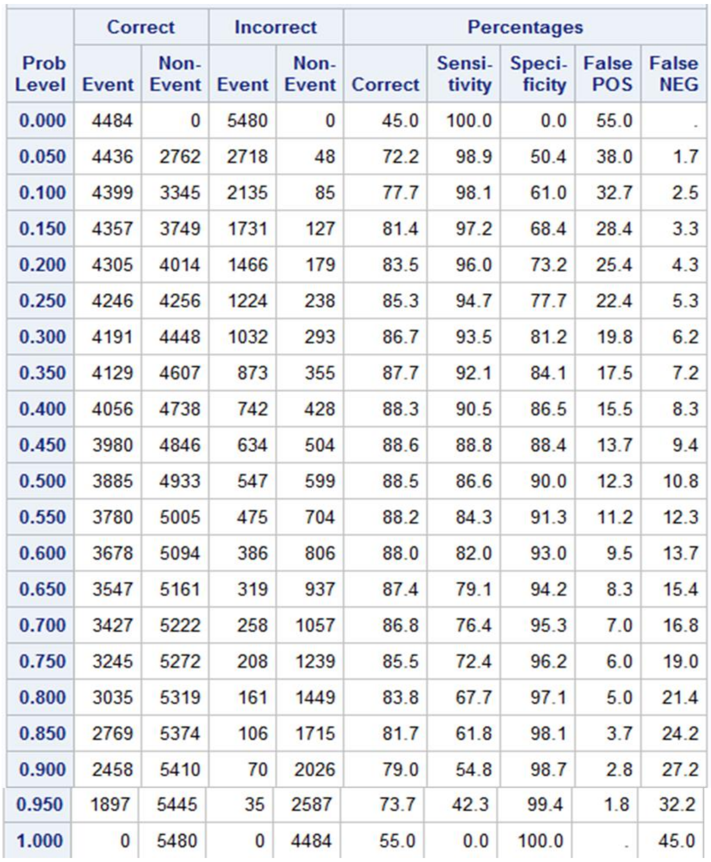
$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{4004}{4484} = 0.8930$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{4004}{4711} = 0.8499$$

$$\text{Fscore} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} = 0.8709$$

2) Logistic Regression

Table 4.3 is the classification table showing the accuracy and True-Positive rate, False-Positive rate, False-Negative rate, etc. for different cutoff probabilities. In order to make the accuracy as high as possible and make the False-Positive rate as low as possible, the cutoff probability for this logistic regression model is determined to be 0.45. The accuracy of this model can achieve 0.886, higher than other classification models.



	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	4484	0	5480	0	45.0	100.0	0.0	55.0	.
0.050	4436	2762	2718	48	72.2	98.9	50.4	38.0	1.7
0.100	4399	3345	2135	85	77.7	98.1	61.0	32.7	2.5
0.150	4357	3749	1731	127	81.4	97.2	68.4	28.4	3.3
0.200	4305	4014	1466	179	83.5	96.0	73.2	25.4	4.3
0.250	4246	4256	1224	238	85.3	94.7	77.7	22.4	5.3
0.300	4191	4448	1032	293	86.7	93.5	81.2	19.8	6.2
0.350	4129	4607	873	355	87.7	92.1	84.1	17.5	7.2
0.400	4056	4738	742	428	88.3	90.5	86.5	15.5	8.3
0.450	3980	4846	634	504	88.6	88.8	88.4	13.7	9.4
0.500	3885	4933	547	599	88.5	86.6	90.0	12.3	10.8
0.550	3780	5005	475	704	88.2	84.3	91.3	11.2	12.3
0.600	3678	5094	386	806	88.0	82.0	93.0	9.5	13.7
0.650	3547	5161	319	937	87.4	79.1	94.2	8.3	15.4
0.700	3427	5222	258	1057	86.8	76.4	95.3	7.0	16.8
0.750	3245	5272	208	1239	85.5	72.4	96.2	6.0	19.0
0.800	3035	5319	161	1449	83.8	67.7	97.1	5.0	21.4
0.850	2769	5374	106	1715	81.7	61.8	98.1	3.7	24.2
0.900	2458	5410	70	2026	79.0	54.8	98.7	2.8	27.2
0.950	1897	5445	35	2587	73.7	42.3	99.4	1.8	32.2
1.000	0	5480	0	4484	55.0	0.0	100.0	.	45.0

Table 4.3

Figure 4.1 is the ROC curve of the model. The Area Under the Curve is larger than 0.95, which indicates the model to be overfitting. One of the possible reasons may be the values of some independent variables (like MoM) only reflect a specific moment in time series.

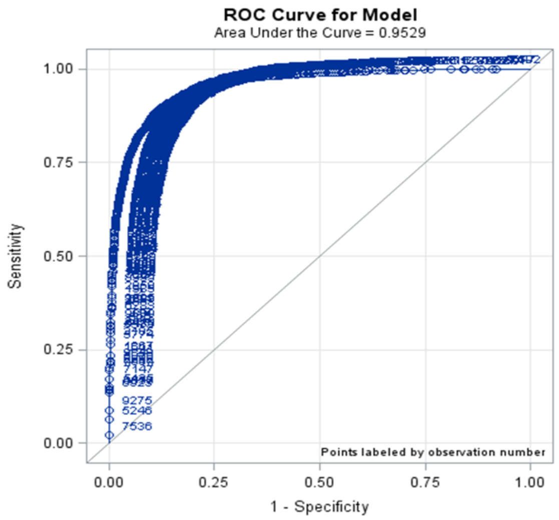
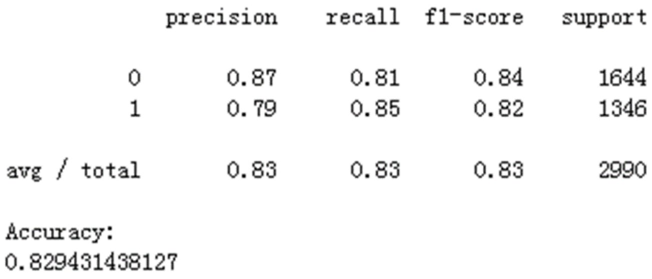


Figure 4.1

3) Naïve Bayes

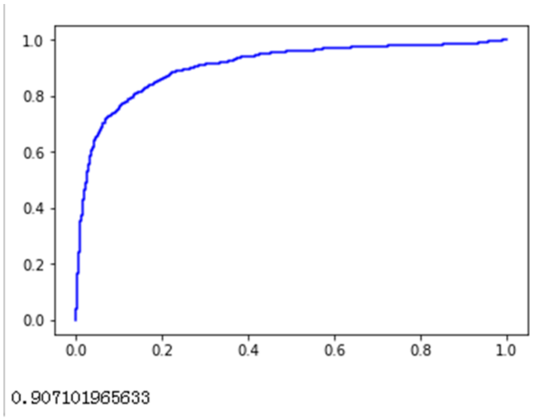
Figure 4.2 shows the performance of the Gaussian Naïve Bayes model. The accuracy and the average precision, recall, f-score is lower than previous two models. One of the possible reasons may be that the features of a record are not independent as assumed.



```
             precision    recall  f1-score   support

          0       0.87      0.81      0.84      1644
          1       0.79      0.85      0.82      1346

avg / total       0.83      0.83      0.83      2990

Accuracy:
0.829431438127
```

Figure 4.2

However, the ROC curve of the Gaussian Naïve Bayes model (Figure 4.3) shows that the Area Under the Curve = 0.9071, indicating a good performance of the model without the overfitting.



```
0.907101965633
```

Figure 4.3

V. Conclusion

1. As the measurement of real estate market, Market Health Index is indicated by the variables including: “SellForGain”, “MoM”, “YoY”, “ForecastYoYPctChange”, “NegativeEquity”, “Delinquency”, “DaysOnMarket”, “Zri”. They can be used for generally evaluate the market health without Market Health Index.
2. Logistic Regression model has better classification accuracy, but is easy to cause overfitting. Naïve Bayes model has relatively poorer accuracy, but less likely to cause the overfitting.