

Графические методы сравнения классификационных моделей

В статье рассмотрены несколько наиболее распространенных графических методов сравнения точности прогнозирования классификационных моделей.

Определения и обозначения

Пусть имеется набор пар независимых одинаково распределенных случайных величин (X_i, Y_i) , где каждая X_i – наблюдаемая, возможно многомерная, состоящая как из дискретных, так и непрерывных компонент случайная величина, а каждая Y_i – ненаблюдаемая бинарная случайная величина, описывающая один из двух классов, к которому принадлежит X_i .

Мы будем называть X_i входной переменной, а Y_i – переменной класса. В случае, если $Y_i = 1$, мы говорим о положительной (positive) классификации X_i , в случае $Y_i = 0$ – об отрицательной (negative). Под X_i можно понимать набор симптомов i -ого пациента, а под Y_i – факт его заболевания.

Каждая классификационная модель определяет измеримую функцию рейтинга (score)

$R: \{X\} \rightarrow [0, 1]$, присваивающую каждой входной переменной X значение рейтинга от 0 до 1.

Большее значение рейтинга $R(X)$ свидетельствует о большей степени уверенности модели, что X принадлежит классу 1, т.е. что $Y=1$. Чаще всего на практике рейтингу соответствует оценка вероятности согласно классификационной модели того, что данный X принадлежит классу 1.

Для простоты будем далее обозначать $R_i := R(X_i)$.

Далее мы предполагаем, что условные вероятности рейтингов $P(R_i \in A | Y_i = 1)$ и

$P(R_i \in A | Y_i = 0)$ имеют функции плотности вероятности относительно меры Лебега f и g соответственно, т.е.

$$P(R_i \in A | Y_i = 1) = \int_A f(x) dx \text{ и}$$

$$P(R_i \in A | Y_i = 0) = \int_A g(x) dx \tag{1}$$

для любого Борелевского множества $A \in B \cap [0, 1]$.

Из (1) следует, что условные относительно результатов классификации математические ожидания измеримых функций от рейтинга $h(R(X))$ для любой Борелевской функции h равны:

$$E(h(R_i) | Y_i = 1) = \frac{1}{P(Y_i = 1)} \int_{\{Y_i=1\}} h(R_i)(\omega) dP(\omega) = \int_0^1 h(x) f(x) dx \text{ и}$$

$$E(h(R_i) | Y_i = 0) = \frac{1}{P(Y_i = 0)} \int_{\{Y_i=0\}} h(R_i)(\omega) dP(\omega) = \int_0^1 h(x) g(x) dx \tag{2}$$

Далее определим F и G как условные относительно результатов классификации функции распределения рейтинга $R(X)$, т.е.

$$F(t) := P(R_i \leq t | Y_i = 1) = \int_0^t f(x) dx \text{ и}$$

$$G(t) := P(R_i \leq t | Y_i = 0) = \int_0^t g(x) dx \quad (3)$$

Для каждой классификационной модели определяется классификатор (classifier) $C : \{X\} \rightarrow \{0,1\}$, приписывающий каждой входной переменной значение ее класса. Чаще всего такой классификатор определяется через некое пороговое значение, приписывающее X к классу 1 в том случае, если рейтинг $R(X)$ больше некоего порога (threshold) t , т.е.

$$C_t(X) = 1 \Leftrightarrow R(X) \geq t \quad (4)$$

Таким образом, имея функцию рейтинга R , мы можем определить посредством формулы (4) семейство классификаторов $\{C_t | t \in [0,1]\}$ для различных значений порога t .

В зависимости от значения Y_i , т.е. фактического значения класса i -ой записи и значения

$C_t(X_i)$, т.е. прогноза значения класса i -ой записи, мы можем получить один из четырех случаев:

1. **True Positive (TP)** – положительный прогноз совпадает с фактическим значением
2. **False Positive (FP)** – положительный прогноз соответствует отрицательному фактическому значению
3. **True Negative (TN)** – отрицательный прогноз совпадает с фактическим значением
4. **False Negative (FN)** – отрицательный прогноз соответствует положительному фактическому значению

Таким образом, для каждого классификатора мы можем получить матрицу сопряженности (confusion matrix), в ячейках которой указывается число записей, соответствующих каждому из описанных случаев:

		Фактическое значение	
		1	0
Прогноз	1	True Positive	False Positive
	0	False Negative	True Negative

Часто вместо абсолютных значений результатов классификации используется их доля в общем числе соответствующих классов:

$$\text{True Postive Rate} := TPR := \frac{TP}{TP + FN}, \quad (5)$$

доля правильно классифицированных положительных классов в общем числе положительных классов.

$$\text{False Postive Rate} := FPR := \frac{FP}{FP + TN}, \quad (6)$$

доля неправильно классифицированных положительных классов в общем числе отрицательных классов.

Так как значения TPR и FPR для одной и той же рейтинговой модели зависят от порога t , определяющего классификатор, то мы иногда будем писать $TPR(t)$ и $FPR(t)$.

По закону больших чисел (5) и (6) являются сильно состоятельными оценками соответствующих условных вероятностей, т.е.

$$TPR(t) \rightarrow P(C_i(X_i) = 1 | Y_i = 1) = P(R(X_i) \geq t | Y_i = 1) = 1 - F(t) = \int_t^1 f(x) dx \quad (7)$$

почти наверняка.

$$FPR(t) \rightarrow P(C_i(X_i) = 1 | Y_i = 0) = P(R(X_i) \geq t | Y_i = 0) = 1 - G(t) = \int_t^1 g(x) dx \quad (8)$$

почти наверняка.

Кривые ROC

Кривые ROC (Receiver Operator Characteristic) предназначены для сравнения качества различных рейтинговых моделей, т.е. того, насколько точно та или иная модель приписывает рейтинги элементам из различных классов. Кривые ROC не предназначены для сравнения классификаторов, т.к. сравнивают рейтинговые модели для всей совокупности классификаторов, определяемых пороговыми значениями.

Кривая ROC для одной рейтинговой модели отображает зависимость доли правильно классифицированных положительных исходов от доли неправильно классифицированных отрицательных исходов для различных значений порогового значения. Т.е. для того, чтобы построить кривую для одной модели, надо для каждого порогового значения отсечения от 0 до 1 определить классификатор и построить по результатам классификации матрицу сопряженности и отобразить на графике точку TPR(t) на оси Y и FPR(t) на оси X.

Модель, соответствующая случайному поиску и служащая для отображения худшей реализации классификационного алгоритма, будет иметь одинаковую долю правильно и неправильно классифицированных положительных и отрицательных случаев соответственно, т.е. каждая точка кривой ROC будет лежать на прямой от точки (0,0) до точки (1,1).

Мы считаем, что любая адекватная классификационная модель должна дать большее относительное число правильных классификаций положительного исхода, чем относительное число неправильных классификаций отрицательных исходов. Это означает, что график TPR(t) от FPR(t) будет находиться выше диагональной кривой (TPR = FPR), соответствующей случайному выбору. Интуитивно, чем лучше модель – тем выше находится график, а идеальным случаем является кривая, соединяющая точки (0,0) с (0,1), а затем (0,1) с (1,1).

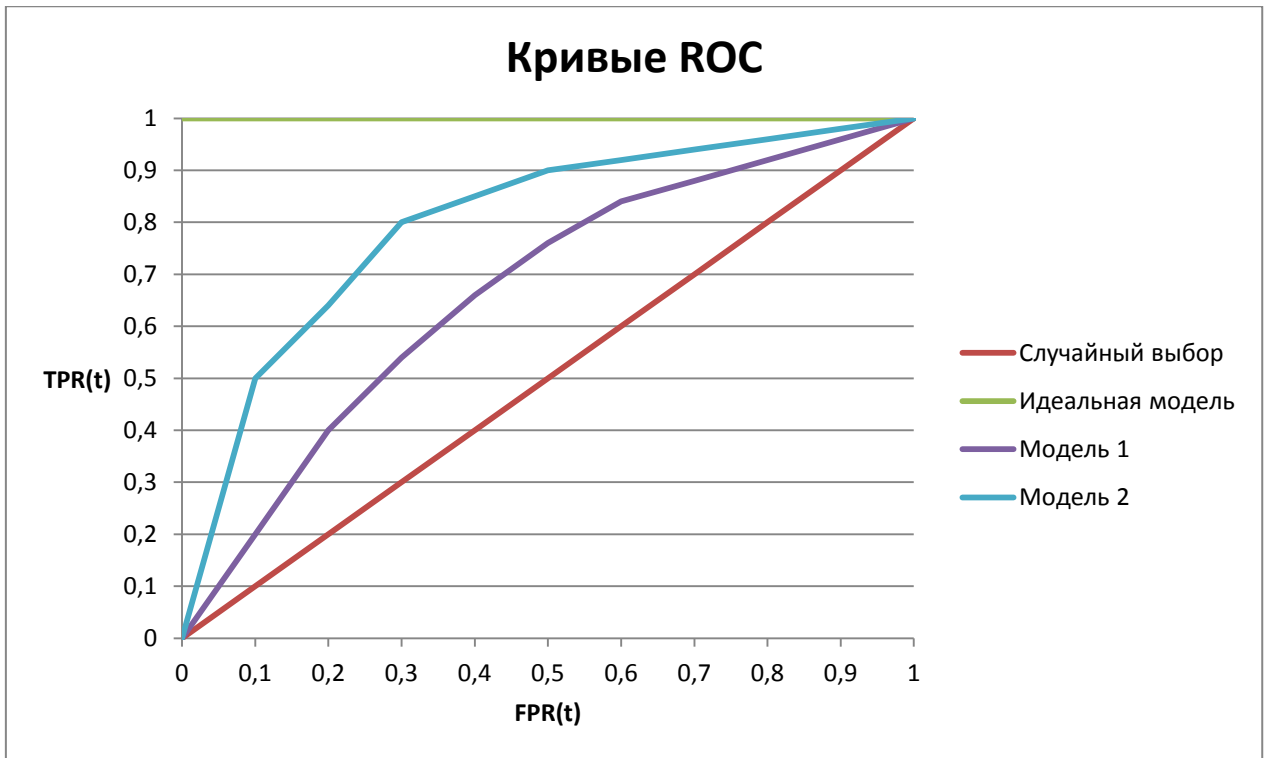


Рис. 1 – Кривые ROC для различных классификационных моделей.

Покажем связь между площадью под графиком кривой ROC (Area Under the Curve – AUC) с качеством выставляемых моделью рейтингов.

Качество рейтинговой модели мы определим как вероятность того, что рейтинг случайно выбранной записи с положительным исходом будет больше рейтинга случайно выбранной записи с отрицательным исходом, т.е. мы случайным образом независимо друг от друга выбираем X_i среди записей с классом 1 и X_j среди записей с классом 0 и сравниваем их рейтинги:

$$RQ := P(R_i > R_j | Y_i = 1, Y_j = 0) \quad (9)$$

Равенство (9) описывает, насколько хорошо модель отличает классы друг от друга.

Получаем:

$$RQ = P(R_i > R_j | Y_i = 1, Y_j = 0) = \frac{P(R_i > R_j, Y_i = 1, Y_j = 0)}{P(Y_i = 1, Y_j = 0)} = \frac{\int \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} dP}{P(Y_i = 1, Y_j = 0)} = \frac{\int \mathbf{1}_{\{r_i > r_j\}} \mathbf{1}_{\{y_i=1\}} \mathbf{1}_{\{y_j=0\}} dP_{(R_i, Y_i, R_j, Y_j)}(r_i, y_i, r_j, y_j)}{P(Y_i = 1, Y_j = 0)} \quad (10)$$

Вследствие независимости (R_i, Y_i) и (R_j, Y_j) имеем $P(Y_i = 1, Y_j = 0) = P(Y_i = 1)P(Y_j = 0)$ и

$$P_{(R_i, Y_i, R_j, Y_j)}(r_i, y_i, r_j, y_j) = P_{(R_i, Y_i)}(r_i, y_i) \otimes P_{(R_j, Y_j)}(r_j, y_j)$$

Следовательно, по теореме Фубини равенство (10) принимает вид:

$$\begin{aligned}
RQ &= \frac{\int \mathbf{1}_{\{y_j=0\}} \left(\int \mathbf{1}_{\{r_i>r_j\}} \mathbf{1}_{\{y_i=1\}} dP_{(R_i, Y_i)}(r_i, y_i) \right) dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_i=1)P(Y_j=0)} = \\
&= \frac{\int \mathbf{1}_{\{y_j=0\}} \left(\int \mathbf{1}_{\{R_i>r_j\}} \mathbf{1}_{\{Y_i=1\}} dP \right) dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_i=1)P(Y_j=0)} = \frac{\int \mathbf{1}_{\{y_j=0\}} P(R_i > r_j, Y_i=1) dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_i=1)P(Y_j=0)} = \\
&= \frac{\int \mathbf{1}_{\{y_j=0\}} \frac{P(R_i > r_j, Y_i=1)}{P(Y_i=1)} dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_j=0)} = \frac{\int \mathbf{1}_{\{y_j=0\}} P(R_i > r_j | Y_i=1) dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_j=0)} = \\
&= \frac{\int \mathbf{1}_{\{y_j=0\}} (1-F(r_j)) dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_j=0)} = 1 - \frac{\int \mathbf{1}_{\{y_j=0\}} F(r_j) dP_{(R_j, Y_j)}(r_j, y_j)}{P(Y_j=0)} = \\
&= 1 - \frac{\int F(R_j) dP}{P(Y_j=0)} = 1 - E(F(R_j) | Y_j=0)
\end{aligned} \tag{11}$$

Учитывая (2), получаем из (11):

$$RQ = 1 - E(F(R_j) | Y_j=0) = 1 - \int_0^1 F(r)g(r)dr = 1 - \int_0^1 F(r)dG(r)$$

Итак,

$$RQ = \int_0^1 (1-F(r))dG(r) \tag{12}$$

Далее рассмотрим площадь под графиком функции $1-F(t)$ от $1-G(t)$, которую мы обозначим как TAUC (Theoretical AUC). Обе функции $1-F(t)$ и $1-G(t)$ – монотонно убывающие, т.к. $F(t)$ и $G(t)$ – монотонно возрастающие функции распределения вероятности. Возьмем произвольное разбиение отрезка $[0,1]$ на отрезки величины не больше Δ вида $[1-G(t_i), 1-G(t_{i+1})]$, где $1=t_1 > t_2 > \dots > t_n = 0$. Площадь под графиком функции $1-F(t)$ от $1-G(t)$ можно записать как предел Римановских сумм:

$$TAUC = \lim_{\Delta \rightarrow 0} \sum_i (1-F(t_i))(1-G(t_{i+1}) - (1-G(t_i))) = \lim_{\Delta \rightarrow 0} \sum_i (1-F(t_i))(G(t_i) - G(t_{i+1})) \tag{13}$$

Так как $[G(t_{i+1}), G(t_i)]$ также разбиение отрезка $[0,1]$ на отрезки величины не больше Δ , то

$$\text{получаем } TAUC = \lim_{\Delta \rightarrow 0} \sum_i (1-F(t_i))(G(t_i) - G(t_{i+1})) = \int_0^1 (1-F(t))dG(t), \text{ а, следовательно, с}$$

учетом (12)

$$TAUC = RQ \tag{14}$$

С другой стороны кривая ROC по определению является графиком функции TPR(t) от FPR(t), а согласно (7) и (8) TPR(t) и FPR(t) – состоятельные оценки функций $1-F(t)$ и $1-G(t)$. Таким образом, можно предположить, что площадь под кривой ROC является состоятельной оценкой площади TAUC под кривой функции $1-F(t)$ от $1-G(t)$ а, следовательно, по (14), оценкой RQ, т.е. AUC \rightarrow TAUC = RQ. Докажем строго, что AUC является оценкой RQ.

По закону больших чисел $\frac{\sum_i 1_{\{Y_i=1\}}}{n} \xrightarrow{P} P(Y=1)$ и $\frac{\sum_i 1_{\{Y_i=0\}}}{n} \xrightarrow{P} P(Y=0)$ как средние значения

последовательностей одинаково распределенных независимых случайных величин.

Следовательно, с учетом независимости Y_i и Y_j при $i \neq j$:

$$\frac{\sum_{i,j} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}}}{n^2} = \frac{\sum_i 1_{\{Y_i=1\}}}{n} \frac{\sum_i 1_{\{Y_i=0\}}}{n} \xrightarrow{P} P(Y_i=1)P(Y_j=0) = P(Y_i=1, Y_j=0); i \neq j. \quad (15)$$

Определим статистику

$$U_n := \frac{1}{n(n-1)} \sum_{i \neq j} 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} \quad (16)$$

$$\text{Очевидно } EU_n = \frac{1}{n(n-1)} \sum_{i \neq j} E 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} = P(R_i > R_j, Y_i=1, Y_j=0) =: \theta \quad (17)$$

$$\text{Далее, } EU_n^2 = \frac{1}{n^2(n-1)^2} \sum_{i \neq j, i' \neq j'} E 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} 1_{\{R_{i'} > R_{j'}\}} 1_{\{Y_{i'}=1\}} 1_{\{Y_{j'}=0\}} \quad (18)$$

Для $i \neq j, i' \neq j'$ возможны следующие варианты:

1. Все индексы попарно различны: $i \neq i', i \neq j', j \neq i', j \neq j'$. Тогда с учетом независимости (R_i, Y_i, R_j, Y_j) и $(R_{i'}, Y_{i'}, R_{j'}, Y_{j'})$ для непересекающихся наборов индексов $\{i, j\}$ и $\{i', j'\}$:

$$\begin{aligned} E 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} 1_{\{R_{i'} > R_{j'}\}} 1_{\{Y_{i'}=1\}} 1_{\{Y_{j'}=0\}} &= \left(E 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} \right) \left(E 1_{\{R_{i'} > R_{j'}\}} 1_{\{Y_{i'}=1\}} 1_{\{Y_{j'}=0\}} \right) = \\ &= P(R_i > R_j, Y_i=1, Y_j=0)^2 = \theta^2 \end{aligned} \quad (19)$$

2. $i \neq i', i \neq j', j \neq i', j = j'$. Тогда индексы $\{i, j, i'\}$ попарно различные, следовательно, случайные величины $(R_i, Y_i), (R_j, Y_j), (R_{i'}, Y_{i'})$ независимы. Интегрируя по частям:

$$\begin{aligned} E 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} 1_{\{R_{i'} > R_{j'}\}} 1_{\{Y_{i'}=1\}} 1_{\{Y_{j'}=0\}} &= E 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} 1_{\{R_{i'} > R_{j'}\}} 1_{\{Y_{i'}=1\}} = \\ &= \int 1_{\{r_i > r_j\}} 1_{\{y_i=1\}} 1_{\{y_j=0\}} 1_{\{r_{i'} > r_{j'}\}} 1_{\{y_{i'}=1\}} dP_{(X_i, R_i)}(x_i, r_i) dP_{(X_j, R_j)}(x_j, r_j) dP_{(X_{i'}, R_{i'})}(x_{i'}, r_{i'}) = \\ &= \int \left(\int 1_{\{r_i > r_j\}} 1_{\{y_i=1\}} dP_{(X_i, R_i)}(x_i, r_i) \right) \left(\int 1_{\{r_{i'} > r_{j'}\}} 1_{\{y_{i'}=1\}} dP_{(X_{i'}, R_{i'})}(x_{i'}, r_{i'}) \right) 1_{\{y_j=0\}} dP_{(X_j, R_j)}(x_j, r_j) = \\ &= \int P(R_i > r_j, Y_i=1) P(R_{i'} > r_j, Y_{i'}=1) 1_{\{y_j=0\}} dP_{(X_j, R_j)}(x_j, r_j) = \\ &= \int P(R_i > r_j, Y_i=1)^2 1_{\{y_j=0\}} dP_{(X_j, R_j)}(x_j, r_j) =: \theta_1 \end{aligned} \quad (20)$$

3. $i = i', i \neq j', j \neq i', j \neq j'$. Тогда индексы $\{i, j, j'\}$ попарно различные, следовательно, случайные величины $(R_i, Y_i), (R_j, Y_j), (R_{j'}, Y_{j'})$ независимы. Интегрируя по частям:

$$\begin{aligned}
& E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_j=0\}} = \\
& = \int \mathbf{1}_{\{r_i > r_j\}} \mathbf{1}_{\{y_i=1\}} \mathbf{1}_{\{y_j=0\}} \mathbf{1}_{\{r_i > r_j\}} \mathbf{1}_{\{y_j=0\}} dP_{(X_i, R_i)}(x_i, r_i) dP_{(X_j, R_j)}(x_j, r_j) dP_{(X_{j'}, R_{j'})}(x_{j'}, r_{j'}) = \\
& = \int \left(\int \mathbf{1}_{\{r_i > r_j\}} \mathbf{1}_{\{y_j=0\}} dP_{(X_j, R_j)}(x_j, r_j) \right) \left(\int \mathbf{1}_{\{r_i > r_j\}} \mathbf{1}_{\{y_j=0\}} dP_{(X_{j'}, R_{j'})}(x_{j'}, r_{j'}) \right) \mathbf{1}_{\{y_i=1\}} dP_{(X_i, R_i)}(x_i, r_i) = \quad (21) \\
& = \int P(R_j < r_i, Y_j = 0) P(R_{j'} < r_i, Y_{j'} = 0) \mathbf{1}_{\{y_i=1\}} dP_{(X_i, R_i)}(x_i, r_i) = \\
& = \int P(R_j < r_i, Y_j = 0)^2 \mathbf{1}_{\{y_i=1\}} dP_{(X_i, R_i)}(x_i, r_i) =: \theta_2
\end{aligned}$$

4. $i \neq i', i \neq j', j = i', j \neq j'$. Получаем:

$$E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_j > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = 0, \quad (22)$$

так как $\mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{Y_j=1\}} = 0$.

5. $i \neq i', i = j', j \neq i', j \neq j'$. Получаем:

$$E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = 0, \quad (23)$$

так как $\mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_i=0\}} = 0$.

6. Случай $i = i', i = j'$ невозможен, так как в этом случае $i' = j'$.

7. Случай $i = i', j = i'$ невозможен, так как в этом случае $i = j$.

8. $i = i', j = j'$. Тогда

$$\begin{aligned}
& E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = \\
& = P(R_i > R_j, Y_i = 1, Y_j = 0) = \theta
\end{aligned} \quad (24)$$

9. $i = j', j = i'$. Тогда

$$E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} = E\mathbf{1}_{\{R_i > R_j\}} \mathbf{1}_{\{Y_i=1\}} \mathbf{1}_{\{Y_j=0\}} \mathbf{1}_{\{R_j > R_i\}} \mathbf{1}_{\{Y_j=1\}} \mathbf{1}_{\{Y_i=0\}} = 0 \quad (25)$$

10. Случай $i = j', j = j'$ невозможен, так как в этом случае $i = j$.

11. Случай $j = i', j = j'$ невозможен, так как в этом случае $i' = j'$.

Из (18), (19)-(25) Получаем:

$$\begin{aligned}
EU_n^2 &= \frac{1}{n^2(n-1)^2} \sum_{i \neq j, i' \neq j'} E1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} 1_{\{R_{i'} > R_{j'}\}} 1_{\{Y_{i'}=1\}} 1_{\{Y_{j'}=0\}} = \\
&= \frac{1}{n^2(n-1)^2} \sum_{i, j, i', j' \text{ попарно различны}} \theta^2 + \frac{1}{n^2(n-1)^2} \sum_{\substack{i, j, i' \text{ попарно различны;} \\ j'=j}} \theta_1 + \\
&+ \frac{1}{n^2(n-1)^2} \sum_{\substack{i, j, j' \text{ попарно различны;} \\ i'=i}} \theta_2 + \frac{1}{n^2(n-1)^2} \sum_{\substack{i, j \text{ попарно различны;} \\ i'=i \text{ и } j'=j}} \theta = \\
&= \frac{n(n-1)(n-2)(n-3)}{n^2(n-1)^2} \theta^2 + \frac{n(n-1)(n-2)}{n^2(n-1)^2} \theta_1 + \\
&+ \frac{n(n-1)(n-2)}{n^2(n-1)^2} \theta_2 + \frac{n(n-1)}{n^2(n-1)^2} \theta
\end{aligned}$$

То есть:

$$EU_n^2 = \frac{(n-2)(n-3)}{n(n-1)} \theta^2 + \frac{(n-2)}{n(n-1)} \theta_1 + \frac{(n-2)}{n(n-1)} \theta_2 + \frac{1}{n(n-1)} \theta$$

Отсюда следует:

$$\begin{aligned}
VarU_n &= EU_n^2 - (EU_n)^2 = \\
&= \frac{(n-2)(n-3)}{n(n-1)} \theta^2 + \frac{(n-2)}{n(n-1)} \theta_1 + \frac{(n-2)}{n(n-1)} \theta_2 + \frac{1}{n(n-1)} \theta - \theta^2 = \\
&= \theta^2 \left(\frac{(n-2)(n-3)}{n(n-1)} - 1 \right) + \frac{(n-2)}{n(n-1)} \theta_1 + \frac{(n-2)}{n(n-1)} \theta_2 + \frac{1}{n(n-1)} \theta = \\
&= \frac{-4n+6}{n(n-1)} \theta^2 + \frac{(n-2)}{n(n-1)} \theta_1 + \frac{(n-2)}{n(n-1)} \theta_2 + \frac{1}{n(n-1)} \theta
\end{aligned} \tag{26}$$

Из (26) следует:

$$nVarU_n = \frac{-4n+6}{n-1} \theta^2 + \frac{n-2}{n-1} \theta_1 + \frac{n-2}{n-1} \theta_2 + \frac{1}{n-1} \theta \rightarrow -4\theta^2 + \theta_1 + \theta_2; (n \rightarrow \infty) \tag{27}$$

Из неравенства Чебышева и (27) следует:

$$\begin{aligned}
P(|U_n - \theta| > \varepsilon) &\leq \frac{VarU_n}{\varepsilon^2} \rightarrow 0; (n \rightarrow \infty) \Rightarrow \\
\Rightarrow U_n &\xrightarrow{P} \theta = P(R_i > R_j, Y_i = 1, Y_j = 0)
\end{aligned} \tag{28}$$

Таким образом, из (28) следует состоятельность статистики U_n . С учетом (15) и (28) получаем:

$$\begin{aligned}
\hat{RQ} &:= \frac{\sum_{i,j} 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}}}{\sum_{i,j} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}}} = \frac{n(n-1)U_n}{\sum_{i,j} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}}} \\
&= \frac{n-1}{n} \frac{U_n}{\sum_{i,j} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}}} \xrightarrow{P} \frac{P(R_i > R_j, Y_i = 1, Y_j = 0)}{P(Y_i = 1, Y_j = 0)} = P(R_i > R_j | Y_i = 1, Y_j = 0) = RQ
\end{aligned} \tag{29}$$

Т.е. мы показали состоятельность статистики \hat{RQ} .

Пусть у нас имеется n_1 записей класса 1 и n_2 записей класса 0. Отсортируем все $n = n_1 + n_2$ записей по значению их рейтингов по возрастанию: $0 \leq R_{j_1} < R_{j_2} < \dots < R_{j_n} \leq 1$ и присвоим каждой записи X_i значение ее ранга r_i в этой отсортированной последовательности. Каждой записи X_i из класса 1 поставим в соответствие ее ранг r_{1i} среди других записей из этого класса. Тогда для записи X_i из класса 1 число записей X_j из класса 0 с меньшим, чем у X_i рейтингом будет равно разности ранга X_i среди всех записей и ранга X_i среди записей из класса 1, т.е.

$$\begin{aligned}
\sum_j 1_{\{R_i > R_j\}} 1_{\{Y_j=0\}} &= \#\{j | Y_j = 0, R_i > R_j\} = r_i - r_{1i}, \forall i: Y_i = 1, \text{ а, следовательно} \\
\sum_{i,j} 1_{\{R_i > R_j\}} 1_{\{Y_i=1\}} 1_{\{Y_j=0\}} &= \sum_{i: Y_i=1} (r_i - r_{1i}) = \sum_{i: Y_i=1} r_i - \sum_{i: Y_i=1} r_{1i} = S_1 - \sum_{i=1}^{n_1} 1 = S_1 - \frac{n_1(n_1+1)}{2},
\end{aligned} \tag{30}$$

где S_1 – сумма всех рейтингов записей из класса 1.

Число всех пар из разных классов равно $n_1 n_2$, таким образом, выражение для \hat{RQ} из (29) принимаем с учетом (30) вид:

$$\hat{RQ} = \frac{\sum_{i: Y_i=1} (r_i - r_{1i})}{n_1 n_2} = \frac{S_1 - \frac{n_1(n_1+1)}{2}}{n_1 n_2} \tag{31}$$

Следует обратить внимание, что оценка (31) аналогична статистике Mann-Whitney-Wilcoxon для двух выборок.

Теперь получим формулу для AUC. Пусть пороговое значение классификатора t уменьшается от максимального значения рейтинга до минимального каждый раз на $R_{j_k} - R_{j_{k-1}}$ за k -ый шаг. Тогда на k -ом шаге уменьшение порогового значения рейтинга приведет к тому, что k -ая запись будет приписана классификатором к классу 1, а соответствующая точка на ROC-графике переместится на $1/n_1$ единиц вверх или на $1/n_2$ единиц вправо в зависимости от того, правильно ли была произведена классификация или нет. Пусть $R_{i_1} < \dots < R_{i_{n_1}}$ отсортированные в порядке возрастания рейтинги записей класса 1. При уменьшении пороговой величины рейтинга на отрезке $[R_{i_{k-1}}, R_{i_k})$, между двумя соседними рейтингами записей класса 1, текущее положение на графике переместится на $1/n_1$ единиц вверх (в классификацию добавилась только одна новая запись класса 1) и на число $1/n_2$, умноженное на количество записей класса 0 с рейтингами в интервале $(R_{i_{k-1}}, R_{i_k})$ вправо. Полученная точка имеет координату по оси X, равную $1/n_2$,

умноженному на число записей класса 0 с рейтингом большим R_{1_k} , т.е. $\frac{1}{n_2} (n_2 - (r_{1_k} - r_{1_{1_k}}))$.

Площадь под графиком FPR(t) от TPR(t) (т.е. функции X от Y) будет равна сумме всех k приращений координаты Y_k (т.е. $1/n_1$), умноженных на координату X_k (т.е. $\frac{1}{n_2} (n_2 - (r_{1_k} - r_{1_{1_k}}))$), а

$$\text{следовательно } \sum_{k=1}^{n_1} \frac{1}{n_1} \frac{1}{n_2} (n_2 - (r_{1_k} - r_{1_{1_k}})) = \frac{1}{n_1 n_2} \left(n_1 n_2 - \sum_{k=1}^{n_1} (r_{1_k} - r_{1_{1_k}}) \right) = 1 - \frac{\sum_{k=1}^{n_1} (r_{1_k} - r_{1_{1_k}})}{n_1 n_2}$$

Так как площадь под графиком TPR(t) от FPR(t) (т.е. функции Y от X) будет равна 1 - площадь под графиком FPR(t) от TPR(t), получаем:

$$AUC = 1 - \left(1 - \frac{\sum_{k=1}^{n_1} (r_{1_k} - r_{1_{1_k}})}{n_1 n_2} \right) = \frac{\sum_{k=1}^{n_1} (r_{1_k} - r_{1_{1_k}})}{n_1 n_2} \quad (32)$$

С учетом (31) получаем:

$$AUC = \hat{RQ} \rightarrow RQ \quad (33)$$

Мы получили, что площадь под графиком кривой ROC является оценкой вероятности того, что рейтинг произвольной записи из положительного класса будет больше рейтинга произвольной записи из отрицательного класса. Таким образом, AUC позволяет сравнивать модели с точки зрения точности выставляемых ими рейтингов.

Lift Chart

Графики Lift Chart также как и ROC предназначены для сравнения моделей с точки зрения точности присваивания рейтинга.

Lift Chart представляет собой график зависимости правильно классифицированных положительных записей TP(t) от всех положительно классифицированных записей TP(t) + FP(t) для всех пороговых значений отсечения рейтинга t. Lift Chart строится следующим образом: для каждой i-ой записи по оси X откладывается отношение ее ранга по убыванию к общему числу

записей (т.е. $\frac{n+1-r_i}{n} = \frac{\{j | R_j \geq R_i\}}{n} = TP(R_i) + FN(R_i)$), а по оси Y – отношение ее ранга по

убыванию среди записей положительного класса к общему числу записей положительного класса

(т.е. $\frac{n_1+1-r_{1_i}}{n_1} = \frac{\{j | R_j \geq R_i, Y_j = 1\}}{n_1} = TP(R_i)$). Таким образом, если перемещаться по индексу

записей в направлении уменьшения рейтинга, то по оси X мы будем перемещаться каждый раз на $1/n$ единиц вправо, а по оси Y – на $1/n_1$ единиц вверх, если соответствующая запись имеет класс 1 и оставаться на предыдущем уровне, если запись имеет класс 0.

В случае случайного присвоения рейтинга записям и последующей сортировки их по значениям рейтингов, записи положительного класса будут равномерно распределены среди всех остальных записей, и поэтому Lift Chart будет представлять собой почти прямую линию из (0,0) в (1,1). В случае идеальной модели, рейтинг каждой положительной записи будет больше рейтинга каждой отрицательной записи, поэтому при уменьшении порогового значения мы будем проходить сначала только через положительные записи, и график будет все время расти; затем, когда положительные записи закончатся, график при прохождении отрицательных записей уже расти не будет. Т.е. Lift Chart для идеальной модели представляет собой соединение двух прямых,

проходящих через точки: $(0,0)$ - $(n_1/n,1)$ - $(1,1)$. График реальной адекватной модели будет находиться между этими двумя экстремальными случаями. В начале оси X он будет расти похоже на график идеальной модели, т.е. расти быстро, т.к. записи с высокими рейтингами будут часто принадлежать классу 1, а в конце – почти не расти, т.к. записи с низким рейтингом редко будут положительными.

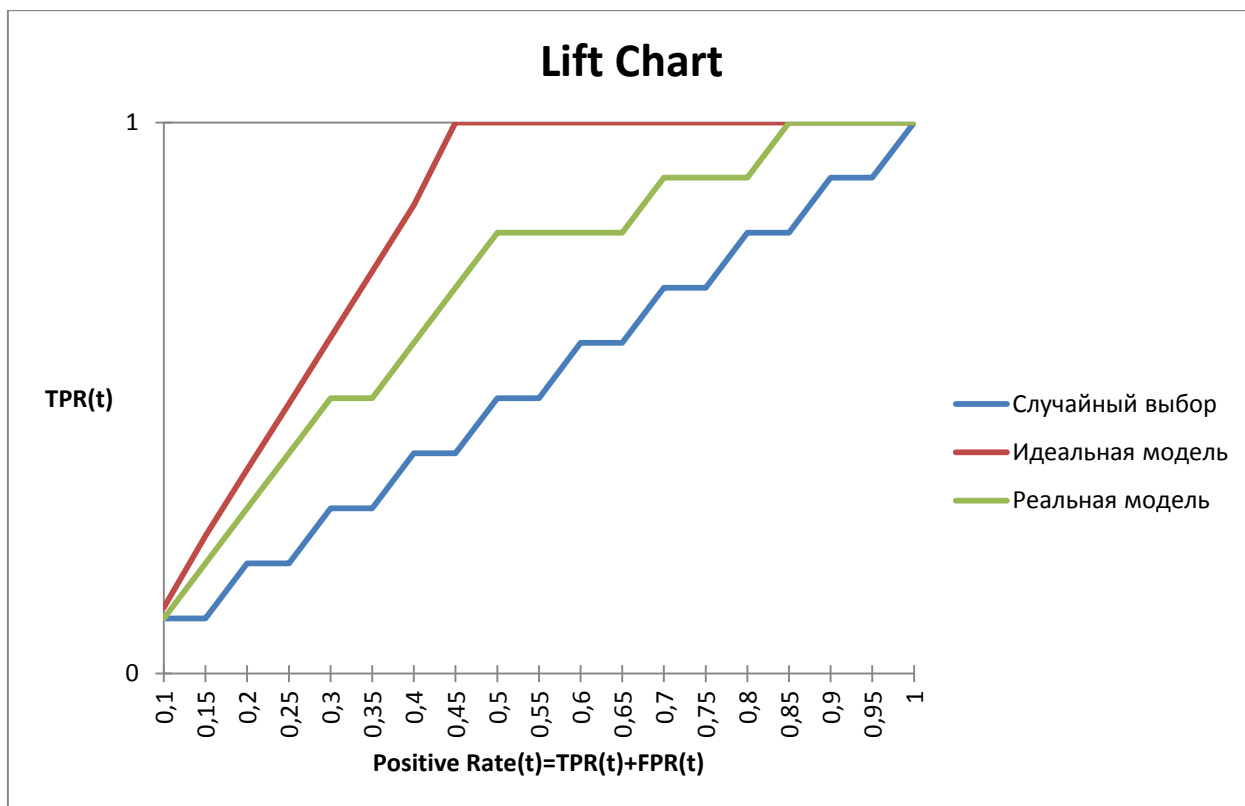


Рис. 2 – Кривые Lift Chart для различных классификационных моделей.

График Lift Chart иллюстрирует, с какой долей записей мы должны «работать», чтобы покрыть заданную долю положительных записей. Так, в случае модели случайного выбора, чтобы покрыть 75% положительных записей мы должны выбрать 75% из всех записей. В случае идеальной модели, чтобы покрыть 75% положительных записей мы должны выбрать 75% среди только первых записей, потому что они будут из положительного класса. В реальной модели, благодаря тому, что записи положительного класса будут в большей степени сосредоточены в начале нашего отсортированного по рейтингу списка записей, чтобы покрыть те же 75% положительных записей нам надо будет работать, например, только с первыми 50% всех записей.

Теперь получим формулу для площади под графиком Lift Chart, которую обозначим как AUC_{LC} .

Пусть $R_{i_1} < \dots < R_{i_{n_1}}$, отсортированные в порядке возрастания рейтинги записей класса 1. При

уменьшении пороговой величины рейтинга на отрезке $[R_{i_{k-1}}, R_{i_k})$, между двумя соседними рейтингами записей класса 1, текущее положение на графике переместится на $1/n_1$ единиц вверх (в классификацию добавилась только одна новая запись класса 1) и на число $1/n$, умноженное на количество записей с рейтингами в интервале $(R_{i_{k-1}}, R_{i_k})$ вправо. Полученная точка имеет координату по оси X, равную $1/n$, умноженному на число записей с рейтингом большим, чем R_{i_k} ,

т.е. $\frac{n+1-r_i}{n}$. Площадь под графиком $FPR(t) + TPR(t)$ от $TPR(t)$ (т.е. функции X от Y или площадь

над графиком Lift Chart) будет равна сумме всех k приращений координаты Y_k (т.е. $1/n_1$),

умноженных на координату X_k , (т.е. $\frac{n+1-r_i}{n}$) минус половину прямоугольника $1/2nn_1$, так как график растет на шаге правильной классификации по диагонали (см. рис. 3):

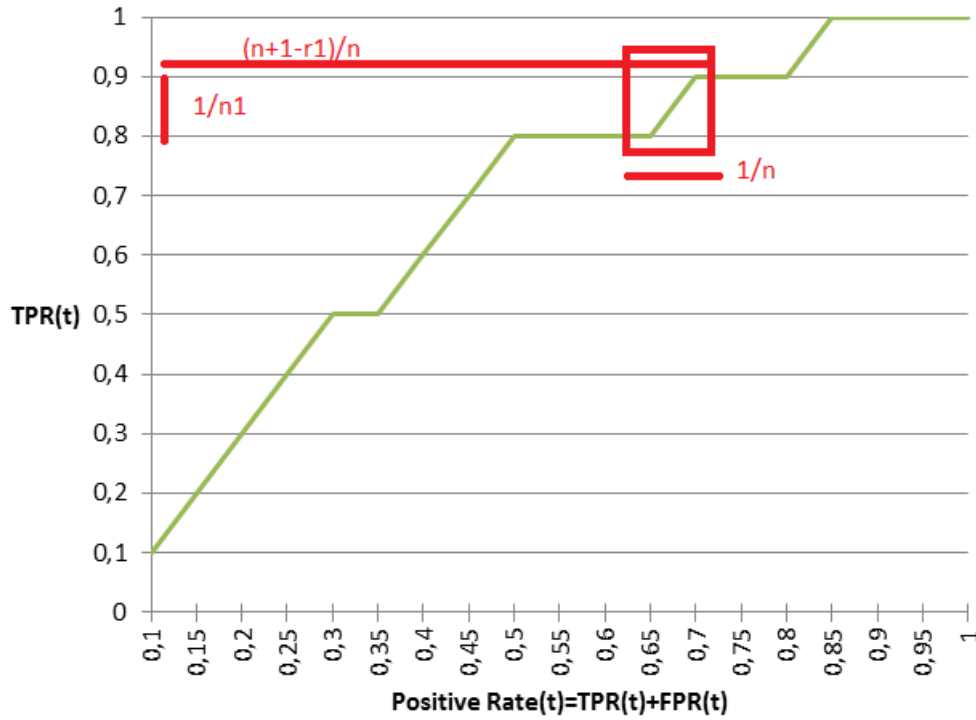


Рис. 3 – Вычисление площади над графиком Lift Chart.

Итак, получаем

$$1 - AUC_{LC} = \sum_{k=1}^{n_1} \frac{1}{n_1} \left(\frac{n+1-r_i}{n} - \frac{1}{2n} \right) = \sum_{k=1}^{n_1} \frac{1}{n_1} \left(1 + \frac{1}{2n} - \frac{r_i}{n} \right) = 1 + \frac{1}{2n} - \sum_{k=1}^{n_1} \frac{r_i}{nn_1}$$

Следовательно,

$$AUC_{LC} = \frac{\sum_{k=1}^{n_1} r_i}{nn_1} - \frac{1}{2n} = \frac{\sum_{k=1}^{n_1} (r_i - 1/2)}{nn_1} \quad (34)$$

Обозначая площадь под ROC-кривой AUC_{ROC} из (31), (33) и (34), получаем:

$$AUC_{ROC} = \frac{\sum_{i,Y_i=1} (r_i - r1_i)}{n_1 n_2} = \frac{\sum_{i,Y_i=1} r_i}{n_1 n_2} - \frac{n_1(n_1+1)}{2n_1 n_2} = \frac{n}{n_2} \left(AUC_{LC} + \frac{1}{2n} \right) - \frac{n_1+1}{2n_2} \Rightarrow$$

$$\Rightarrow AUC_{LC} = \frac{n_2}{n} AUC_{ROC} - \frac{1}{2n} + \frac{n_1+1}{2n} = \frac{n_2}{n} AUC_{ROC} + \frac{n_1}{2n}$$

Следовательно:

$$AUC_{LC} = \frac{n_2}{n} AUC_{ROC} + \frac{n_1}{2n} \quad (35)$$

Таким образом, площадь под графиком Lift Chart является линейной функцией от площади под соответствующей кривой ROC, и, следовательно, является линейной функцией от оценки

вероятности того, что произвольная запись класса 1 имеет больший рейтинг, чем произвольная запись класса 0. Т.е. площадь под Lift Chart также является показателем качества выставления рейтинга моделью.

Profit Chart

Profit Chart предназначена для определения оптимальной величины порогового значения функции классификатора. Так, если на первом шаге при помощи кривых ROC или Lift Chart мы получаем лучшую с точки зрения точности выставления рейтинга модель, то на втором этапе при помощи Profit Chart мы получаем для отобранной модели лучшее (по определенному критерию) пороговое значения для функции классификатора.

Для того, чтобы получить Profit Chart (диаграмма роста прибыли) сначала надо составить матрицу вознаграждений и штрафов для каждого варианта фактического и прогнозируемого значения. Так, в случае диагностики онкологических заболеваний, «польза» от правильного диагноза рака больше «пользы» от правильного диагноза отсутствия рака. Аналогично, «вред» ошибки от постановки диагноза «здоров» раковому больному больше «вреда» постановки ракового диагноза здоровому человеку. Другой случай – рассылка полиграфической рекламы потенциальным клиентам. Случай TP соответствует прибыли в случае покупки минус затраты на рассылку, FP – штраф в сумме затрат на рассылку, TN – ноль, FN – ноль. Таким образом, матрицы вознаграждений и штрафов для различных практических случаев могут сильно отличаться друг от друга в зависимости от специфики классификационной задачи. Задача определения величин вознаграждений и штрафов лежит в предметной области и решается экспертами.

Profit Chart строится следующим образом: для определенной матрицы вознаграждений и штрафов и для выбранной рейтинговой модели строится график, где по оси X отображаются значения порога отсека, а по оси Y – величина суммарной прибыли для этого порога. Величина прибыли вычисляется как сумма вознаграждений или штрафов для каждой записи в зависимости от результата классификации и фактического значения класса этой записи. Далее, на полученном графике ищется пороговое значение, соответствующее максимуму прибыли. В дальнейшем, посредством этого порогового значения определяется классификатор, служащий для прогнозирования класса записей.

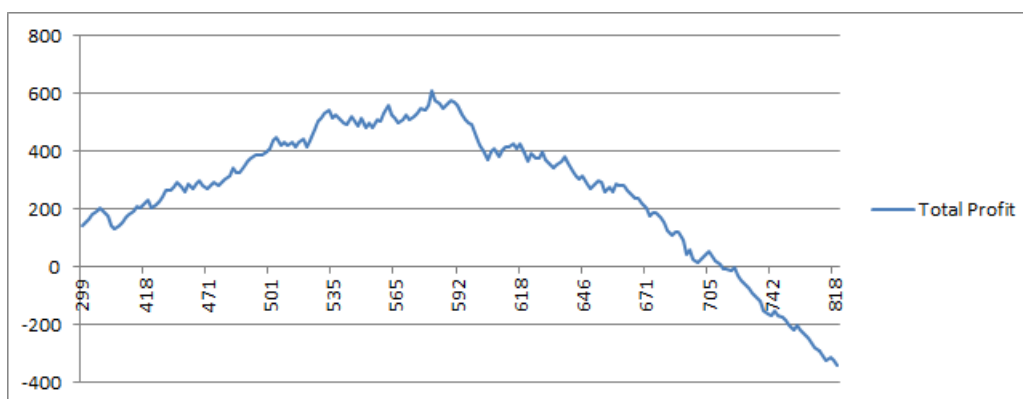


Рис. 4 – Profit Chart

Матрица вознаграждений и штрафов:

False Positive Cost	10
False Negative Cost	5
True Positive Profit	10
True Negative Profit	3

Оптимальное пороговое значение: 581

Accuracy Chart

Графики Accuracy Chart предназначены для сравнения точности определенного класса классификационных моделей. При построении Accuracy Chart мы исходим из того, что в качестве рейтинга записи модель использует вероятность того, что запись принадлежит классу 1, а в качестве порогового значения используется величина 50%. Выбор такого порога соответствует решающей функции в задаче проверки гипотез $H_0=\{Y=1\}$ и $H_1=\{Y=0\}$, минимизирующей вероятность принятия ошибочной гипотезы. Действительно, пусть $\hat{Y}(X)$ - оценка класса Y на основании данных X . Получим вероятность правильной оценки:

$$P(\hat{Y}(X) = Y | X) = \sum_{i=0}^1 P(\hat{Y}(X) = i, Y = i | X) \quad (36)$$

Значение класса Y оказывает влияние на распределение только X , а оценка \hat{Y} осуществляется только на основании X . То есть, при фиксированном X оценка \hat{Y} уже не зависит от Y (см. рис. 5).

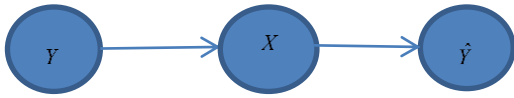


Рис. 5 – Байесовская сеть, описывающая оценку класса.

Следовательно, $P(Y, \hat{Y} | X) = P(\hat{Y} | X)P(Y | X)$

$$\text{С учетом (22) получаем: } P(\hat{Y}(X) = Y | X) = \sum_{i=0}^1 P(\hat{Y}(X) = i | X)P(Y = i | X) \quad (37)$$

Так как $\hat{Y}(X)$ - измеримая функция от X , то $P(\hat{Y}(X) = i | X) = E(1_{\{\hat{Y}(X)=i\}} | X) = 1_{\{\hat{Y}(X)=i\}}$ и, следовательно

$$P(\hat{Y}(X) = Y | X) = \sum_{i=0}^1 P(Y = i | X) 1_{\{\hat{Y}(X)=i\}} \quad (38)$$

Определим оценку класса следующим образом:

$$\hat{Y}(X) := \begin{cases} 1, P(Y = 1 | X) \geq P(Y = 0 | X) = 1 - P(Y = 1 | X) \\ 0, P(Y = 1 | X) < P(Y = 0 | X) = 1 - P(Y = 1 | X) \end{cases} = \begin{cases} 1, P(Y = 1 | X) \geq 0,5 \\ 0, P(Y = 1 | X) < 0,5 \end{cases} \quad (39)$$

Пусть $\omega \in \Omega$ - произвольный элемент вероятностного пространства и

$s \in \{0, 1\} : P(Y = s | X)(\omega) = \max \{P(Y = j | X)(\omega) | j \in \{0, 1\}\}$, тогда из определения (39) следует $\hat{Y}(\omega) = s$. Пусть $\varphi(X)$ - любая другая оценка Y , $\varphi(X) = k \in \{0, 1\}$, тогда, с учетом (38) получаем:

$$\begin{aligned} P(\hat{Y}(X) = Y | X)(\omega) &= \sum_{i=0}^1 P(Y = i | X)(\omega) \underbrace{1_{\{\hat{Y}(X)=i\}}(\omega)}_{=1 \Leftrightarrow i=s} = P(Y = s | X)(\omega) \geq \\ &\geq P(Y = k | X)(\omega) = P(Y = k | X)(\omega) 1_{\{\varphi(X)=k\}}(\omega) = \sum_{i=0}^1 P(Y = i | X)(\omega) \underbrace{1_{\{\varphi(X)=i\}}(\omega)}_{=1 \Leftrightarrow i=k} = \\ &= P(\varphi(X) = Y | X)(\omega) \end{aligned}$$

$$\text{Т.е. } P(\hat{Y}(X) = Y | X)(\omega) \geq P(\varphi(X) = Y | X)(\omega), \forall \omega \in \Omega \quad (40)$$

и, следовательно,

$$P(\hat{Y}(X) = Y) = \int \underbrace{P(\hat{Y}(X) = Y | X)}_{\geq P(\varphi(X) = Y | X)}(\omega) dP(\omega) \geq \int P(\varphi(X) = Y | X)(\omega) dP(\omega) = P(\varphi(X) = Y) \quad (41)$$

Неравенство (41) означает, что вероятность правильной классификации для классификатора, определенного посредством (39) – максимальная среди всех классификаторов (оценок Y по X), или, что вероятность ошибки классификатора (39) – минимальная.

$$P(\hat{Y}(X) \neq Y) \leq P(\varphi(X) \neq Y) \quad (42)$$

Другими словами, из (39) и (42) следует, что в случае, если в качестве функции рейтинга модели выбрана вероятность принадлежности строки положительному классу, а пороговым значением – 0,5, то вероятность ошибки при такой классификации будет минимальной для множества всех оценок класса Y по X .

Accuracy Chart строится для таких моделей следующим образом: записи сортируются по вероятности, с которой модель осуществляет прогноз для обоих значения класса 0 и 1. По оси X откладывается ранг записи по убыванию, деленный на число записей, а по оси Y – график растет на $1/n$, если прогноз осуществлен правильно и остается на предыдущем уровне, если была осуществлена ошибка. Т.е. график вырастет только на относительное число правильно классифицированных записей

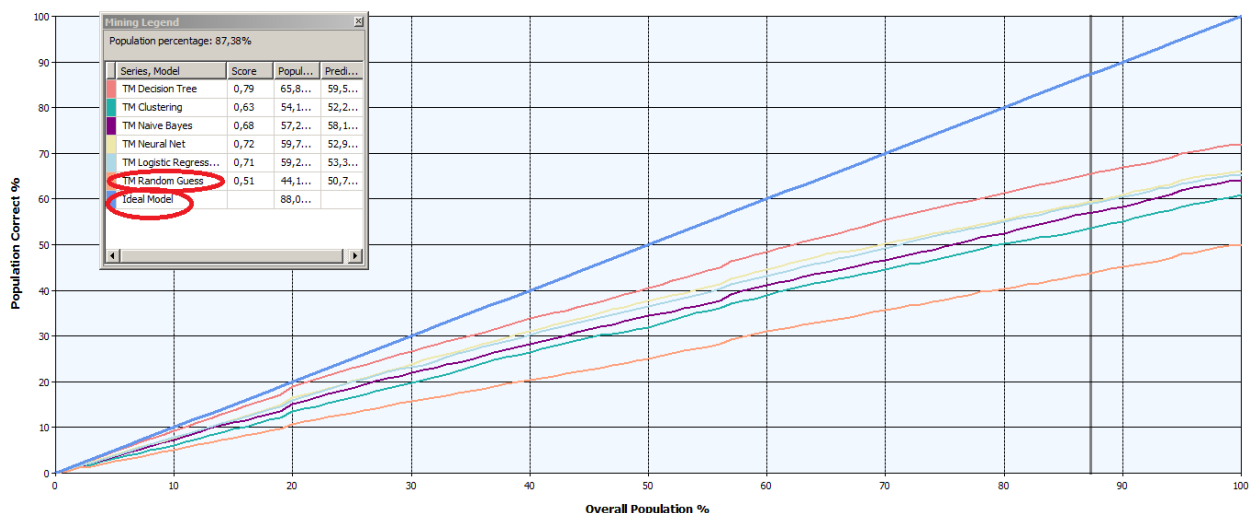


Рис. 6 – Accuracy Chart.

График идеальной модели идет наверх по диагонали из точки (0,0) в точку (1,1), так как все записи классифицируются правильно. Случайный выбор идет также по диагонали из (0,0) к точке, соответствующей априорной вероятности наиболее вероятного исхода (если априорные вероятности исходов, например 40% и 60% для 1 и 0 соответственно, то прогноз по порогу 50% даст исход 0 во всех случаях, что будет соответствовать правильному угадыванию в 60% случаев). График реальной адекватной модели будет находиться между этими двумя экстремальными случаями. Его вид будет вогнутым (рост вначале больше, чем в конце оси X), так как записи в начале координат будут иметь большую вероятность прогноза, а, следовательно, большее их число на самом деле будут иметь тот же класс, а в конце оси X вероятность прогноза будет мала, и меньшее число записей будут правильно классифицированы.

Максим Гончаров
Ноябрь 2010

maxim.goncharov@spellabs.ru
maxgon@microsoft.com