

Contents

Statistical concepts used:.....	2
Chi-square test:.....	2
Naïve Bayes Model:.....	2
Question 1 - Are the active and non-active customers randomly distributed by the type of payment method that they use?	4
CODE:	4
OUTPUT:.....	5
Interpretation:	5
Question 2 - Does customer churn depend upon the gender of the client?	6
CODE:	6
OUTPUT:.....	7
Interpretation:	7
Question 3 - Is there a relationship between customer churn and type of contract?	8
CODE:	8
OUTPUT:.....	9
Interpretation:	9
Question 4 - What is the conditional probability of customer churn given a particular contract type and payment method?	10
CODE:	10
OUTPUT:.....	11
Interpretation:	11

Statistical concepts used:

Chi-square test:

The Chi-square test of association evaluates relationships between categorical variables. Like any statistical hypothesis test, the Chi-square test has both a null hypothesis and an alternative hypothesis.

- Null hypothesis: There are no relationships between the categorical variables. If you know the value of one variable, it does not help you predict the value of another variable.
- Alternative hypothesis: There are relationships between the categorical variables. Knowing the value of one variable *does* help you predict the value of another variable.

The Chi-square test of independence works by comparing the distribution that you observe to the distribution that you expect if there is no relationship between the categorical variables. In the Chi-square context, the word “expected” is equivalent to what you would expect if the null hypothesis were true. If your observed distribution is sufficiently different than the expected distribution (no relationship), you can reject the null hypothesis and infer that the variables are related.

For a Chi-square test, a p-value that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. You can conclude that a relationship exists between the categorical variables.

Link: <https://statisticsbyjim.com/hypothesis-testing/chi-square-test-independence-example/>

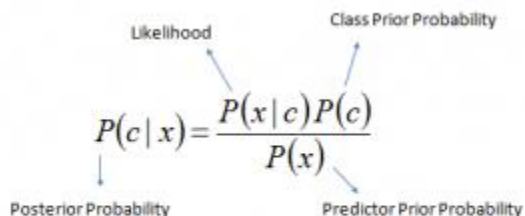
Naïve Bayes Model:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, some fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:



The diagram shows the equation $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with arrows pointing from labels to parts of the equation: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Above,

- $P(c|x)$ is the posterior probability of *class (c, target)* given *predictor (x, attributes)*.

- $P(c)$ is the prior probability of *class*.
- $P(x/c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Link: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

Question 1 - Are the active and non-active customers randomly distributed by the type of payment method that they use?

CODE:

```
WITH
Observed AS (
    SELECT
        [PaymentMethod],
        COUNT(*) AS [Total],
        SUM(1-[stopped]) AS [Active],
        SUM([stopped]) AS [NonActive]
    FROM (SELECT *, IIF([Churn] = 'Yes', 1, 0) as [stopped]
        FROM [dbo].[Telco Churn]) pm
    GROUP BY [PaymentMethod]
),
Rate AS (
    SELECT
        Observed.[PaymentMethod],
        Observed.[Active] * 1.0 / Observed.[Total] * 100 AS [Active Rate],
        Observed.[NonActive] * 1.0 / Observed.[Total] * 100 AS [Non Active Rate]
    FROM Observed
),
Total AS (
    SELECT
        'Total' AS [PaymentMethod],
        SUM(Observed.[Total]) AS [Total],
        SUM(Observed.[Active]) AS [Active],
        SUM(Observed.[NonActive]) AS [NonActive],
        SUM(Observed.[Active]) * 1.0 / SUM(Observed.[Total]) * 100 AS [Active Rate],
        SUM(Observed.[NonActive]) * 1.0 / SUM(Observed.[Total]) * 100 AS [Non Active Rate],
        NULL AS [Active Expected],
        NULL AS [Non Active Expected],
        NULL AS [Active Chi],
        NULL AS [Non Active Chi]
    FROM Observed
),
Expected AS (
    SELECT
        Observed.[PaymentMethod],
        Total.[Active Rate] * Observed.[Total] / 100 AS [Active Expected],
        Total.[Non Active Rate] * Observed.[NonActive] / 100 AS [Non Active Expected],
        POWER((Observed.[Active] - (Total.[Active Rate] * Observed.[Total] / 100)), 2)/(Total.[Active Rate]
* Observed.[Total] / 100) AS [Active Chi],
        POWER((Observed.[NonActive] - (Total.[Non Active Rate] * Observed.[Total] / 100)), 2)/(Total.[Non
Active Rate] * Observed.[Total] / 100) AS [Non Active Chi]
    FROM Observed, Total
)
SELECT
    Observed.*,
    Rate.[Active Rate], Rate.[Non Active Rate],
    Expected.[Active Expected], Expected.[Non Active Expected], Expected.[Active Chi], Expected.[Non
Active Chi]
FROM Observed
JOIN Rate ON Observed.[PaymentMethod] = Rate.[PaymentMethod]
JOIN Expected ON Observed.[PaymentMethod] = Expected.[PaymentMethod]
UNION
SELECT * FROM Total
ORDER BY Observed.[PaymentMethod];
GO
```

OUTPUT:

	Total	Active	NonActive	ActiveRate	Non Active Rate	Active Expected	Non Active Expected	Active Deviation	Non Active Deviation	Active Chi	Non Active Chi
Bank transfer (automatic)	1544	1286	258	83.29015544	16.70984456	1134.268919	68.46542666	151.7310805	189.5345733	20.297056	56.188856
Credit card (automatic)	1522	1290	232	84.75689882	15.24310118	1118.107057	61.56581002	171.8929433	170.43419	26.426077	73.155979
Electronic check	2365	1294	1071	54.71458774	45.28541226	1737.400256	284.2111316	-443.4002556	786.7888684	113.159754	313.263012
Mailed check	1612	1304	308	80.89330025	19.10669975	1184.223768	81.7339202	119.7762317	226.2660798	12.114556	33.537033
Total	7043	5174	1869	73.46301292	26.53698708					171.997443	476.14488
Degree of freedom	3										
p-value	3.6824E-140										

Interpretation:

Based on the output generated and p-value of 3.6824E-140, we can conclude that payment method and customer churn have a significant relationship. Therefore, the payment method feature should be part of future models that will be used to predict customer churn.

Question 2 - Does customer churn depend upon the gender of the client?

CODE:

```
WITH
Observed AS (
    SELECT
        [gender],
        COUNT(*) AS [Total],
        SUM(1-[stopped]) AS [Active],
        SUM([stopped]) AS [NonActive]
    FROM (SELECT *, IIF([Churn] = 'Yes', 1, 0) as [stopped]
        FROM [dbo].[Telco Churn]) g
    GROUP BY [gender]
),
Rate AS (
    SELECT
        Observed.[gender],
        Observed.[Active] * 1.0 / Observed.[Total] * 100 AS [Active Rate],
        Observed.[NonActive] * 1.0 / Observed.[Total] * 100 AS [Non Active Rate]
    FROM Observed
),
Total AS (
    SELECT
        'Total' AS [gender],
        SUM(Observed.[Total]) AS [Total],
        SUM(Observed.[Active]) AS [Active],
        SUM(Observed.[NonActive]) AS [NonActive],
        SUM(Observed.[Active]) * 1.0 / SUM(Observed.[Total]) * 100 AS [Active Rate],
        SUM(Observed.[NonActive]) * 1.0 / SUM(Observed.[Total]) * 100 AS [Non Active Rate],
        NULL AS [Active Expected],
        NULL AS [Non Active Expected],
        NULL AS [Active Chi],
        NULL AS [Non Active Chi]
    FROM Observed
),
Expected AS (
    SELECT
        Observed.[gender],
        Total.[Active Rate] * Observed.[Total] / 100 AS [Active Expected],
        Total.[Non Active Rate] * Observed.[NonActive] / 100 AS [Non Active Expected],
        POWER((Observed.[Active] - (Total.[Active Rate] * Observed.[Total] / 100)), 2) / (Total.[Active Rate]
        * Observed.[Total] / 100) AS [Active Chi],
        POWER((Observed.[NonActive] - (Total.[Non Active Rate] * Observed.[Total] / 100)), 2) / (Total.[Non
        Active Rate] * Observed.[Total] / 100) AS [Non Active Chi]
    FROM Observed, Total
)
SELECT
    Observed.*,
    Rate.[Active Rate], Rate.[Non Active Rate],
    Expected.[Active Expected], Expected.[Non Active Expected], Expected.[Active Chi], Expected.[Non
    Active Chi]
FROM Observed
JOIN Rate ON Observed.[gender] = Rate.[gender]
JOIN Expected ON Observed.[gender] = Expected.[gender]
UNION
SELECT * FROM Total
ORDER BY Observed.[gender];
GO
```

OUTPUT:

	Total	Active	NonActive	ActiveRate	Non Active Rate	Active Expected	Non Active Expected	Active Deviation	Non Active Deviation	Active Chi	Non Active Chi
Female	3488	2549	939	73.07912844	26.92087156	2562.389891	249.1823087	-13.38989067	689.8176913	0.069969	0.193698
Male	3555	2625	930	73.83966245	26.16033755	2611.610109	246.7939798	13.38989067	683.2060202	0.06865	0.190047
Total	7043	5174	1869	73.46301292	26.53698708					0.138619	0.383745
Degree of freedom	1										
p-value	0.47										

Interpretation:

Based on the output generated and p-value of 0.47, we can conclude that customer gender and customer churn have no significant relationship. Therefore, the gender feature should not be included in future models that will be used to predict customer churn.

Question 3 - Is there a relationship between customer churn and type of contract?

CODE:

```
WITH
Observed AS (
    SELECT
        [Contract],
        COUNT(*) AS [Total],
        SUM(1-[stopped]) AS [Active],
        SUM([stopped]) AS [NonActive]
    FROM (SELECT *, IIF([Churn] = 'Yes', 1, 0) as [stopped]
        FROM [dbo].[Telco Churn]) c
    GROUP BY [Contract]
),
Rate AS (
    SELECT
        Observed.[Contract],
        Observed.[Active] * 1.0 / Observed.[Total] * 100 AS [Active Rate],
        Observed.[NonActive] * 1.0 / Observed.[Total] * 100 AS [Non Active Rate]
    FROM Observed
),
Total AS (
    SELECT
        'Total' AS [Contract],
        SUM(Observed.[Total]) AS [Total],
        SUM(Observed.[Active]) AS [Active],
        SUM(Observed.[NonActive]) AS [NonActive],
        SUM(Observed.[Active]) * 1.0 / SUM(Observed.[Total]) * 100 AS [Active Rate],
        SUM(Observed.[NonActive]) * 1.0 / SUM(Observed.[Total]) * 100 AS [Non Active Rate],
        NULL AS [Active Expected],
        NULL AS [Non Active Expected],
        NULL AS [Active Chi],
        NULL AS [Non Active Chi]
    FROM Observed
),
Expected AS (
    SELECT
        Observed.[Contract],
        Total.[Active Rate] * Observed.[Total] / 100 AS [Active Expected],
        Total.[Non Active Rate] * Observed.[NonActive] / 100 AS [Non Active Expected],
        POWER((Observed.[Active] - (Total.[Active Rate] * Observed.[Total] / 100)), 2) / (Total.[Active
Rate] * Observed.[Total] / 100) AS [Active Chi],
        POWER((Observed.[NonActive] - (Total.[Non Active Rate] * Observed.[Total] / 100)), 2) / (Total.[Non
Active Rate] * Observed.[Total] / 100) AS [Non Active Chi]
    FROM Observed, Total
)
SELECT
    Observed.*,
    Rate.[Active Rate], Rate.[Non Active Rate],
    Expected.[Active Expected], Expected.[Non Active Expected], Expected.[Active Chi], Expected.[Non
Active Chi]
FROM Observed
JOIN Rate ON Observed.[Contract] = Rate.[Contract]
JOIN Expected ON Observed.[Contract] = Expected.[Contract]
UNION
SELECT * FROM Total
ORDER BY Observed.[Total];
GO
```


OUTPUT:

	Total	Active	NonActive	Active Rate	Non Active Rate	Active Expected	Non Active Expected	Active Deviation	Non Active Deviation	Active Chi	Non Active Chi
One year	1473	1307	166	88.73048201	11.26951799	1082.11018	44.05139855	224.8898197	121.9486014	46.737783	129.385388
Two year	1695	1647	48	97.16814159	2.831858407	1245.198069	12.7377538	401.801931	35.2622462	129.653904	358.924185
Month-to-month	3875	2220	1655	57.29032258	42.70967742	2846.691751	439.1871362	-626.6917507	1215.812864	137.964551	381.930759
Total	7043	5174	1869	73.46301292	26.53698708					314.356238	870.240332
Degree of freedom	2										
p-value	5.863E-258										

Interpretation:

The p-value generated above indicates a significant relationship between customer churn and type of contract chosen at the time of onboarding. We should include this feature when trying to predict customer churn.

Question 4 - What is the probability of customer churn given a particular contract type and payment method?

CODE:

```
WITH
ContractDim AS (
    SELECT
        [Contract],
        AVG(IIF([Churn] = 'Yes', 1.0, 0)) AS prob
    FROM [dbo].[Telco Churn]
    GROUP BY [Contract]
),
PaymentMethodDim AS (
    SELECT
        [PaymentMethod],
        AVG(IIF([Churn] = 'Yes', 1.0, 0)) AS prob
    FROM [dbo].[Telco Churn]
    GROUP BY [PaymentMethod]
),
Overall AS (
    SELECT
        AVG(IIF([Churn] = 'Yes', 1.0, 0)) AS prob
    FROM [dbo].[Telco Churn]
),
Actual AS (
    SELECT
        [Contract],
        [PaymentMethod],
        AVG(IIF([Churn] = 'Yes', 1.0, 0)) AS prob
    FROM [dbo].[Telco Churn]
    GROUP BY [Contract], [PaymentMethod]
)
SELECT
    [Contract],
    [Contract Probability],
    [PaymentMethod],
    [PaymentMethod Probability],
    [Predicted Probability],
    [Actual Probability]
FROM (
    SELECT
        ContractDim.[Contract],
        ContractDim.[prob] AS [Contract Probability],
        PaymentMethodDim.[PaymentMethod],
        PaymentMethodDim.[prob] AS [PaymentMethod Probability],
        POWER(Overall.[prob], -1) * ContractDim.[prob] * PaymentMethodDim.[prob] AS [Predicted Probability],
        Actual.prob AS [Actual Probability]
    FROM ContractDim
    CROSS JOIN PaymentMethodDim
    CROSS JOIN Overall
    JOIN Actual
    ON ContractDim.[Contract] = Actual.[Contract]
    AND PaymentMethodDim.[PaymentMethod] = Actual.[PaymentMethod]
) dim
ORDER BY [Contract], [PaymentMethod];

GO
```

OUTPUT:

Contract	Contract Probability	Payment Method	Payment Method Probability	Predicted Probability	Actual Probability	Actual - Predicted
Month-to-month	0.427096	Bank transfer (automatic)	0.167098	0.268935	0.341256	7.23%
Month-to-month	0.427096	Credit card (automatic)	0.152431	0.245329	0.327808	8.25%
Month-to-month	0.427096	Electronic check	0.452854	0.728842	0.537297	-19.15%
Month-to-month	0.427096	Mailed check	0.191066	0.30751	0.315789	0.83%
One year	0.112695	Bank transfer (automatic)	0.167098	0.070962	0.097186	2.62%
One year	0.112695	Credit card (automatic)	0.152431	0.064733	0.103015	3.83%
One year	0.112695	Electronic check	0.452854	0.192315	0.184438	-0.79%
One year	0.112695	Mailed check	0.191066	0.081141	0.068249	-1.29%
Two year	0.028318	Bank transfer (automatic)	0.167098	0.017831	0.033687	1.59%
Two year	0.028318	Credit card (automatic)	0.152431	0.016266	0.022375	0.61%
Two year	0.028318	Electronic check	0.452854	0.048325	0.07738	2.91%
Two year	0.028318	Mailed check	0.191066	0.020389	0.007853	-1.25%

Interpretation:

The Naïve-Bayes model using the two significant features determined in question 1 & 3 is quite useful and can predict probability of customer churn with good precision across most combinations of contract type and payment method. However, when the contract type is 'month-to-month' and payment method is 'electronic check', the predicted probability of churn is off by 19.15% from the actual probability which is highest error in prediction in our model. We can reduce this error by testing and including other significant features in our model.