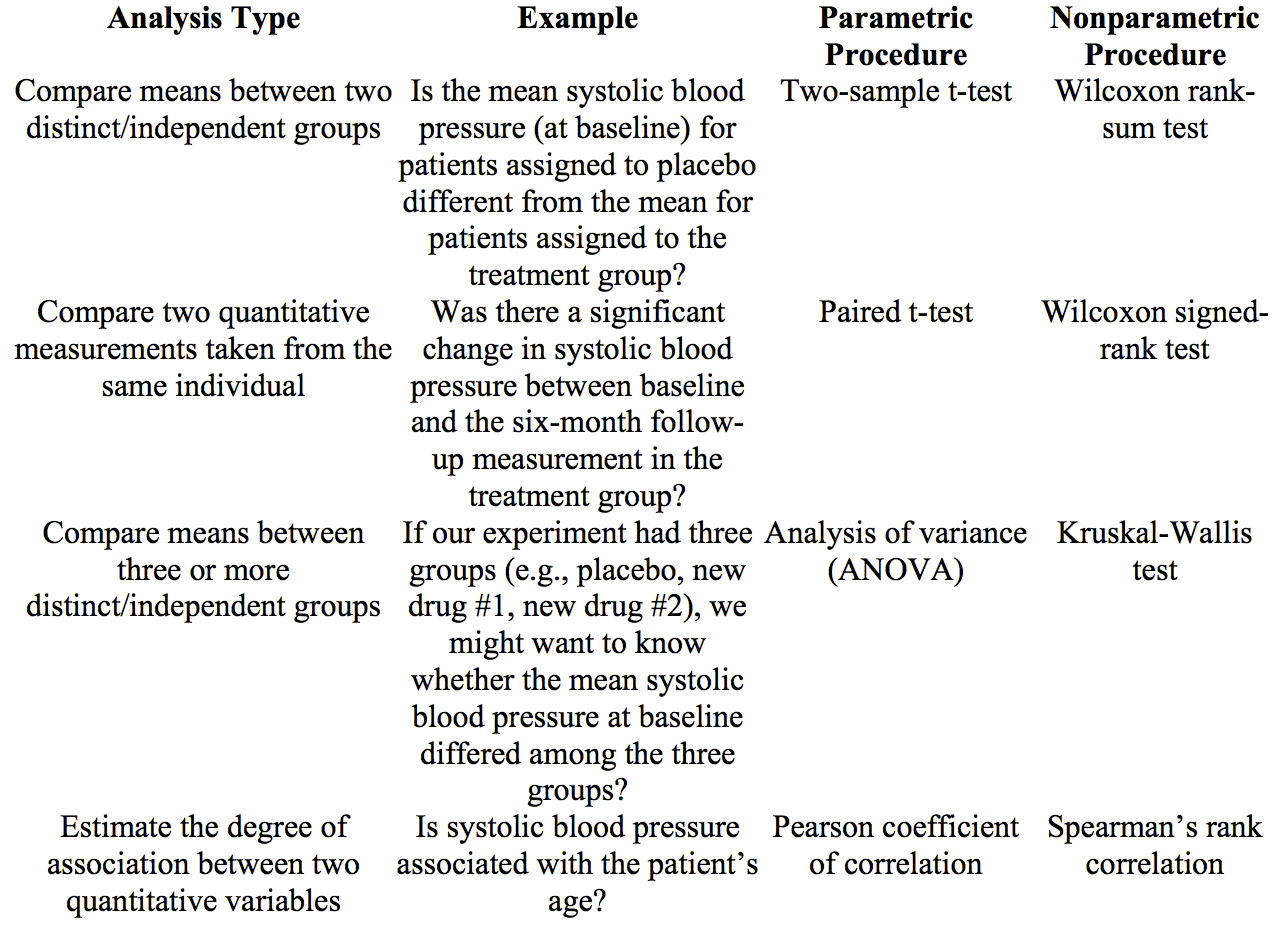
<https://mathbitsnotebook.com/Algebra1/StatisticsData/STShapes.html> 怎么读boxplot

<https://github.com/Duhaichuan/Nonparametric-Projects>

这个的链接展示的code其实也不错的

宝藏网站：<https://www.statstest.com/mann-whitney-u-test/>

也是不错的：<https://stats.idre.ucla.edu/stata/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-stata/>



还有我知乎的收藏的非参数也不错很好对应着看

Jonckheere's trend test

In [statistics](https://en.wikipedia.org/wiki/Statistics), the **Jonckheere trend test**[[1]](https://en.wikipedia.org/wiki/Jonckheere%27s_trend_test#cite_note-jonck-1) (sometimes called the **Jonckheere–Terpstra**[[2]](https://en.wikipedia.org/wiki/Jonckheere%27s_trend_test#cite_note-2)**test**) is a test for an ordered [alternative hypothesis](https://en.wikipedia.org/wiki/Alternative_hypothesis) within an independent samples (between-participants) design. It is similar to the [Kruskal–Wallis test](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_test) in that the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) is that several independent samples are from the same population.(所以也要满足这个assumptions of an identically shaped and scaled distribution了) However, with the Kruskal–Wallis test there is no a priori ordering of the populations from which the samples are drawn. When there is an *a priori* ordering, the Jonckheere test has more [statistical power](https://en.wikipedia.org/wiki/Statistical_power) than the Kruskal–Wallis test. The test was developed by [A. R. Jonckheere](https://en.wikipedia.org/wiki/Aimable_Robert_Jonckheere), who was a psychologist and statistician at [University College London](https://en.wikipedia.org/wiki/University_College_London).

The null and alternative hypotheses can be conveniently expressed in terms of population medians for *k* populations (where *k* > 2). Letting *θi* be the population [median](https://en.wikipedia.org/wiki/Median) for the *i*th population, the null hypothesis is:

Kruskal–Wallis one-way analysis of variance

{\displaystyle H\_{0}:\theta \_{1}=\theta \_{2}=\cdots =\theta \_{k}}

The **Kruskal–Wallis test** by ranks, **Kruskal–Wallis *H* test**[[1]](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance#cite_note-Laerd-1) (named after [William Kruskal](https://en.wikipedia.org/wiki/William_Kruskal) and [W. Allen Wallis](https://en.wikipedia.org/wiki/W._Allen_Wallis)), or **one-way ANOVA on ranks**[[1]](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance#cite_note-Laerd-1) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method for testing whether samples originate from the same distribution.[[2]](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance#cite_note-2)[[3]](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance#cite_note-3)[[4]](https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance#cite_note-4) It is used for comparing two or more independent samples of equal or different sample sizes. It extends the [Mann–Whitney *U* test](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test), which is used for comparing only two groups. The parametric equivalent of the Kruskal–Wallis test is the [one-way analysis of variance](https://en.wikipedia.org/wiki/One_way_anova) (ANOVA). 所以说他的assumption和rank sum 差不多的

Since it is a non-parametric method, the Kruskal–Wallis test does not assume a [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

the test is generally considered to be robust to ties.

Permutational analysis of variance (F test)

**Permutational multivariate analysis of variance** (**PERMANOVA**),[[1]](https://en.wikipedia.org/wiki/Permutational_analysis_of_variance#cite_note-1) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) [multivariate](https://en.wikipedia.org/wiki/Multivariate_statistics) [statistical test](https://en.wikipedia.org/wiki/Statistical_test). PERMANOVA is used to compare groups of objects and test the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) that the [centroids](https://en.wikipedia.org/wiki/Centroids) and [dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion) of the groups as defined by measure space are equivalent for all groups. A rejection of the null hypothesis means that either the centroid and/or the spread of the objects is different between the groups. Hence the test is based on the prior calculation of the distance between any two objects included to the experiment. PERMANOVA shares some resemblance to [ANOVA](https://en.wikipedia.org/wiki/ANOVA) where they both measure the [sum-of-squares](https://en.wikipedia.org/wiki/Analysis_of_variance#Logic) within and between group and make use of [F test](https://en.wikipedia.org/wiki/F_test) to compare within-group to between-group variance. However, while ANOVA bases the [significance](https://en.wikipedia.org/wiki/Statistical_significance) of the result on assumption of normality, PERMANOVA draws tests for significance by comparing the actual F test result to that gained from random permutations of the objects between the groups. Moreover, whilst PERMANOVA tests for similarity based on a chosen distance measure, ANOVA tests for similarity of the group [averages](https://en.wikipedia.org/wiki/Averages).

In the traditional analysis of variance (ANOVA) F-test, we are testing the null hypothesis of equality (no differences) in treatment effects H0 : τ1 = τ2 = · · · = τk against the alternative hypothesis H1 : τi 6= τj for some i 6= j. That is, if H1 is true, then all treatment effects are not equal.

However, if the assumptions are violated, (that is, the residual errors are not normally distributed with constant variance), then a permutation F-test may be appropriate.

# Sign test

From Wikipedia, the free encyclopedia

[Jump to navigation](https://en.wikipedia.org/wiki/Sign_test#mw-head)[Jump to search](https://en.wikipedia.org/wiki/Sign_test#searchInput)

The **sign test** is a statistical method to test for consistent differences between pairs of observations, such as the weight of subjects before and after treatment. Given pairs of observations (such as weight pre- and post-treatment) for each subject, the sign test determines if one member of the pair (such as pre-treatment) tends to be greater than (or less than) the other member of the pair (such as post-treatment)

**Assumptions:**

* **Data distribution:** The Sign test is a non–parametric (distribution free) test, so we do not assume that the data is normally distributed.
* **Two sample:** Data should be from two samples.  The population may differ for the two samples.
* **Dependent sample:** Dependent samples should be a paired sample or matched. Also known as ‘before–after’ sample.

**Types of sign test:**

1. **One sample:**We set up the hypothesis so that + and – signs are the values of random variables having equal size.
2. **Paired sample:** This test is also called an alternative to the [paired sample t-test](http://www.statisticssolutions.com/resources/directory-of-statistical-analyses/paired-sample-t-test).  This test uses the + and – signs in paired sample tests or in before-after study. In this test, null hypothesis is set up so that the sign of + and – are of equal size, or the population means are equal to the sample mean.

### **Paired Data**

Two data sets are "paired" when the following one-to-one relationship exists between values in the two data sets.

* Each data set has the same number of data points.
* Each data point in one data set is related to one, and only one, data point in the other data set.

An example of paired data would be a before-after drug test. The researcher might record the blood pressure of each subject in the study, before and after a drug is administered. These measurements would be paired data, since each "before" measure is related only to the "after" measure from the same subject.

**Wilcoxon rank-sum test**

<https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/>

https://www.statstest.com/mann-whitney-u-test/

# 这个讲的太清楚了,慢慢读一句句读好啊

 is a [nonparametric](https://en.wikipedia.org/wiki/Nonparametric_statistics) [test](https://en.wikipedia.org/wiki/Statistical_hypothesis_test) of the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) that the probability that a randomly selected value from one population is less than a randomly selected value from a second population is equal to the probability of being greater.

This test can be used to investigate whether two *independent* samples were selected from populations having the same distribution. The Mann-Whitney *U* test is often used when the assumptions of the independent samples *t*-test are violated.[[1]](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#cite_note-Pearce-1) A similar nonparametric test used on *dependent* samples is the [Wilcoxon signed-rank test](https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test).

A very general formulation is to assume that:

1. The responses are [ordinal](https://en.wikipedia.org/wiki/Ordinal_measurement) (i.e., one can at least say, of any two observations, which is the greater),
2. Under the null hypothesis H0, the distributions of both populations are equal.[[4]](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#cite_note-4)
3. The alternative hypothesis *H*1 is that the distributions are not equal.
4. 1. The two samples are independent of one another  
   2. The two populations have equal variance or spread （euqal shape也就是same distribution, 注意这不代表是normal, normal dis是由mean ,variance来定义的， 而我我们的非参数rank sum 的是由variance 和 median 来决定的，他们可以是same distri 但是不是normal,看链接有图，还有你想啊我们的Ho假设他们是same dist, 所以如果拒绝的话那肯定是移动了，这也是为马这个是detect for location shift）

# Wilcoxon signed-rank test

<https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php>

# 这个讲的太清楚了,慢慢读一句句读好啊

From Wikipedia, the free encyclopedia

[Jump to navigation](https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test#mw-head)[Jump to search](https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test#p-search)

The **Wilcoxon signed-rank test** is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) [statistical hypothesis test](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing) used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a [paired difference test](https://en.wikipedia.org/wiki/Paired_difference_test)). It can be used as an alternative to the [paired Student's t-test](https://en.wikipedia.org/wiki/Student%27s_t-test) (also known as "*t*-test for matched pairs" or "*t*-test for dependent samples") when the distribution of the difference between two samples' means cannot be assumed to be [normally distributed](https://en.wikipedia.org/wiki/Normally_distributed).[[1]](https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test#cite_note-1) A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution.

1. Your **dependent variable** should be measured at the **ordinal** or **continuous （kg）level**. Examples of **ordinal variables(好，很好， 极好)**
2. Data are paired and come from the same population. Your **independent variable** should consist of **two categorical**, **"related groups"** or **"matched pairs"**. "Related groups" indicates that the same subjects are present in both groups. The reason that it is possible to have the same subjects in each group is because each subject has been measured on two occasions on the same dependent variable. For example, you might have measured 10 individuals' performance in a spelling test (the dependent variable) before and after they underwent a new form of computerized teaching method to improve spelling. Y
3. datas are distributed symmetrically around the median(**symmetrical in shape**
4. Each pair is chosen randomly and independently[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)

## Assumptions

The Wilcoxon signed rank test does not assume that the data are sampled from a Gaussian distribution. However it does assume that the datas are distributed symmetrically around the median(**symmetrical in shape**.). If the distribution is asymmetrical, the P value will not tell you much about whether the median is different than the hypothetical value.

Like all statistical tests, the Wilcoxon signed rank test assumes that the errors are [independent](https://www.graphpad.com/guides/prism/7/statistics/the_need_for_independent_samples.htm). The term “error” refers to the difference between each value and the group median. The results of a Wilcoxon test only make sense when the scatter is random – that any factor that causes a value to be too high or too low affects only that one value.

Kolmogorov–Smirnov test

<https://www.youtube.com/watch?v=ZO2RmSkXK3c>

<https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>

In [statistics](https://en.wikipedia.org/wiki/Statistics), the **Kolmogorov–Smirnov test** (**K–S test** or **KS test**) is a [nonparametric test](https://en.wikipedia.org/wiki/Nonparametric_statistics) of the equality of continuous (or discontinuous, see [Section 2.2](https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test#Discrete_and_mixed_null_distribution)), one-dimensional [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) that can be used to compare a [sample](https://en.wikipedia.org/wiki/Random_sample) with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). It is named after [Andrey Kolmogorov](https://en.wikipedia.org/wiki/Andrey_Kolmogorov) and [Nikolai Smirnov](https://en.wikipedia.org/wiki/Nikolai_Smirnov_(mathematician)).

<https://www.graphpad.com/guides/prism/7/statistics/interpreting_results_kolmogorov-smirnov_test.htm>

## **Ansari-Bradley Test**

Performs the Ansari-Bradley two-sample test for a difference in scale parameters.

Between sample independence are mutually independent

Continuity assumption: both F1 and F2 are continuous distributions

Location assumption: θ1 = θ2 (median)

have the same probability distribution but that their common distribution is not specified. Formally stated, this null hypothesis is

*H*0 : [*F*(*t*) = *G*(*t*), for every *t*].The typical alternative hypothesis in a two-sample dispersion problem specifies that the *Y* population has greater (or less) variability associated with it than does the *X* population.

参考书上5.1

The Difference Between a Chi-Square Test and a McNemar Test

<https://www.theanalysisfactor.com/difference-between-chi-square-test-and-mcnemar-test/>

Fisher's exact test is more accurate than the [chi-square test](http://www.biostathandbook.com/chiind.html) or [G–test](http://www.biostathandbook.com/gtestind.html)of independence when the expected numbers are small. I recommend you use Fisher's exact test when the total sample size is less than 1000, and use the chi-square or G–test for larger sample sizes. See the [web page on small sample sizes](http://www.biostathandbook.com/small.html) for further discussion of what it means to be "small".

**Fisher's Exact test** is a way to **test** the association **between** two categorical variables when you have small cell sizes (expected values less than 5). **Chi**-**square test** is used when the cell sizes are expected to be large

A screenshot of a cell phone

Description automatically generated  **Mann–Whitney *U* test**

貌似错了？

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

<https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests>

Hi Francesco (弗朗车斯科),

it doesn't matter about salary. I'm interested in what I can do, what I can learn from Stocknack.

What I am thinking about is like I can be volunteer work for two or three months, see what I can do for your company, then you and I will decided leave or not.

问题：

It’s goanna be part time job right? How many hour should I work?

What are my main tasks for this job?

While, you have mention that there is one more data scientist , I am wondering, How many years of experience has he /she had in this field? Do we work together? , or we will have different tasks. I mean I don’t have a lot real industry experience, Sometimes we do need team work, need people who has experience in this field to solve problems together, or to lead us to get the job done.

With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results.

Abstract— With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results

Abstract— With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results

Abstract— With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results

Abstract— With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results

With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results.

Abstract— With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results

Abstract— With the rapid development of communication

technology, the field of telecommunication faces complex

challenges due to the number of vibrant competitive service

providers. Customer Churn is the major issue that faces by the

Telecommunication industries in the world. Churn is the activity

of customers leaving the company and discarding the services

offered by it, due to the dissatisfaction with the services. The

main areas of this research contend with the ability to identify

potential churn customers, cluster customers with similar

consumption behavior and mine the relevant patterns embedded

in the collected data. The primary data collected from customers

were used to create a predictive churn model that obtain

customer churn rate of five telecommunication companies. For

model building, classified the relevant variables with the use of

the Pearson chi-square test, cluster analysis, and association rule

mining. Using the Weka, the cluster results produced the

involvement of customers, interest areas and reasons for the

churn decision to enhance marketing and promotional activities.

Using the Rapid miner, the association rule mining with the FP-

Growth component was expressed rules to identify

interestingness patterns and trends in the collected data have a

huge influence on the revenues and growth of the

telecommunication companies. Then, the C5.0 Decision tree

algorithm tree, the Bayesian Network algorithm, the Logistic

Regression algorithm, and the Neural Network algorithms were

developed using the IBM SPSS Modeler 18. Finally, comparative

evaluation is performed to discover the optimal model and test

the model with accurate, consistent and reliable results