UNIT III

CHAPTER 3

Association Rules and Regression

Syllabus Topics

Advanced Analytical Theory and Methods: Association Rules- Overview, a-priori algorithm, evaluation of candidate rules, case study-transactions in grocery store, validation and testing, diagnostics. Regression- linear, logistics, reasons to choose and cautions, additional regression models.

Syllabus Topic : Advanced Analytical Theory and Methods : Association Rules - Overview

3.1 Advanced Analytical Theory and Methods : Association Rules-Overview

Q. 3.1.1 Explain Association rules. (Refer section 3.1) **(4 Marks)**

- Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.
- It is intended to identify strong rules discovered in databases using some measures of interestingness.
- Based on the concept of strong rules, Rakesh Agrawal, Tomasz Imieliński and Arun Swami introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets.
- For example, the rule {onions,potatoes}=>{burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat.
- Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.
- In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, continuous production, and bioinformatics.
- In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.
- The problem of association rule mining is defined as :
  - Let $I = \{i_1,i_2,....,i_n\}$ be a set of n binary attributes called items.
  - Let $D = \{t_1,t_2,....,t_m\}$ be a set of transactions called the database.
- Each transaction in D has a unique transaction ID and contains a subset of the items in I.
- A rule is defined as an implication of the form :

  $X => Y$ where $X,Y \subseteq I$
- In Agrawal, Imieliński, Swami segement a rule is defined only between a set and a single item

  $X => i_j$ for $i_j \in I$
- Every rule is composed by two different sets of items, also known as itemsets X and Y, where X is called antecedent or left-hand-side (LHS) and Y is called as consequent or right-hand-side (RHS).

- To illustrate the concepts, we use a small example from the supermarket domain. The set of items is I = {milk, bread, butter} and in the table is shown a small database containing the items, where, in each entry, the value 1 means the presence of the item in the corresponding transaction, and the value 0 represents the absence of an item in that transaction.
- An example rule for the supermarket could be {butter, bread}=>{milk} meaning that if butter and bread are bought, customers also buy milk.
- This example is extremely small.
- In practical applications, a rule needs a support of several hundred transactions before it can be considered statistically significant, and datasets often contain thousands or millions of transactions.

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter
1	1	1	0
2	0	0	1
3	0	0	0
4	1	1	1
5	0	1	0

- Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time.
- Association rule generation is usually split up into two separate steps:
- A minimum support threshold is applied to find all *frequent itemsets* in a database.
- A minimum confidence constraint is applied to these frequent itemsets in order to form rules.
- While the second step is straightforward, the first step needs more attention.
- Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations).
- The set of possible itemsets is the power set over I and has size $2^n - 1$ (excluding the empty set which is not a valid itemset).
- Although the size of the power-set grows exponentially in the number of items n in I , efficient search is possible using the downward-closure property of support (also called anti-monotonicity).
- This will guarantee that for a frequent itemset, all its subsets are also frequent and thus no infrequent itemset can be a subset of a frequent itemset.
- Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent itemsets.

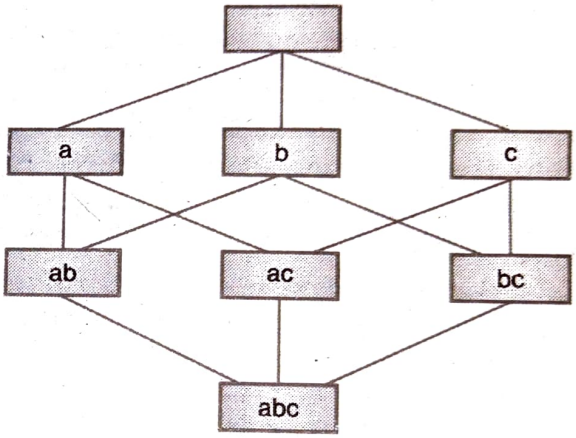

Fig. 3.1.1 : Frequent itemset lattice

- In frequent itemset lattice, the color of the box indicates how many transactions contain the combination of items.
- Note that lower levels of the lattice can contain at most the minimum number of their parents' items; e.g. {ac} can have only at most min(a,c) items. This is called the *downward-closure property*.

Syllabus Topic : Apriori Algorithm

3.2 Apriori Algorithm

Q. 3.2.1 Explain in detail Apriori algorithm with example. (Refer section 3.2) **(8 Marks)**

- Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.
- It proceeds by identifying the frequent individual items in the database and extending them to larger and larger

item sets as long as those item sets appear sufficiently often in the database.

- The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.
- The Apriori algorithm was proposed by Agrawal and Srikant in 1994.
- Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).
- Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).
- Each transaction is seen as a set of items (an itemset). Given a threshold C, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database.
- Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data.
- The algorithm terminates when no further successful extensions are found.
- Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length K from item sets of length K-1.
- Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent K-length item sets.
- After that, it scans the transaction database to determine frequent item sets among the candidates.
- The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of $\in$. Usual set theoretic notation is employed, note that T is a multiset. C_k is the candidate set for level k.
- At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, observing the downward closure lemma.
- count[c] accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero.
- Usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

Apriori(T,$\in$)

$L_1 \leftarrow$ {large 1 – itemsets}

$k \leftarrow 2$

while $L_{k-1} \neq 0$

$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\}$

$- \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \notin L_{k-1}\}$

for transactions $t \in T$

$C_t \{c \mid c \in C_k \wedge c \subseteq t\}$

for candidates $c \in C_t$

count(c) $\leftarrow$ count(c) + 1

$L \{c \mid c \in C_k \wedge \text{count}(c) \geq 6\}$

$k \leftarrow k + 1$

return $\cup L_k$

Examples

Example 1

- Consider the following database, where each row is a transaction and each cell is an individual item of the transaction :

alpha	beta	epsilon
alpha	beta	theta
alpha	beta	epsilon
alpha	beta	theta

- The association rules that can be determined from this database are the following :
  1. 100% of sets with alpha also contain beta
  2. 50% of sets with alpha, beta also have epsilon
  3. 50% of sets with alpha, beta also have theta

- We can also illustrate this through a variety of examples.

Example 2

- Assume that a large supermarket tracks sales data by stock-keeping unit (SKU) for each item: each item such as "butter" or "bread" is identified by a numerical SKU.
- The supermarket has a database of transactions where each transaction is a set of SKUs that were bought together.
- Let the database of transactions consist of following itemsets :

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

- We will use Apriori to determine the frequent item sets of this database. To do this, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value 3 is the *support threshold.*
- The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately. By scanning the database for the first time, we obtain the following result

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

- All the itemsets of size 1 have a support of at least 3, so they are all frequent.
- The next step is to generate a list of all pairs of the frequent items.
- For example, regarding the pair {1,2}: the first table of Example 2 shows items 1 and 2 appearing together in three of the itemsets; therefore, we say item {1,2} has support of three.

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

- The pairs {1,2}, {2,3}, {2,4}, and {3,4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {1,3} and {1,4} are not. Now, because {1,3} and {1,4} are not frequent, any larger set which contains {1,3} or {1,4} cannot be frequent. In this way, we can *prune* sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:

Item	Support
{2,3,4}	2

- In the example, there are no frequent triplets. {2,3,4} is below the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.
- We have thus determined the frequent sets of items in the database, and illustrated how some items were not counted because one of their subsets was already known to be below the threshold.

3.2.1 Limitations of Apriori Algorithm

Q. 3.2.2 What are the limitations of Apriori Algorithm ? (Refer section 3.2.1) **(2 Marks)**

- Apriori, while historically significant, suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms.
- Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan).
- Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all $2^{|s|} - 1$ of its proper subsets.

Syllabus Topic : Evaluation of Candidate Rules

3.3 Evaluation of Candidate Rules

Q. 3.3.1 How to evaluation Candidate rules? (Refer section 3.3) **(4 Marks)**

- The itemsets which we have seen in the previous section can form candidate rules such as X implies Y (X → Y).
- ***Confidence*** can be defined as the measure of assurance or trustworthiness linked with all the discovered rules. Mathematically, confidence can be considered as the percent of transactions which include both X and Y out of all the transactions that contain X.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support } (X \wedge Y)}{\text{Support } (X)}$$

- For example, if {bread,butter,milk} has a support of 0.20 and {bread,butter} has same support, the confidence of rule {bread,butter} → {milk} is 1, which indicates that hundred percent of the time a customer buys bread and butter will buy milk as well.
- Hence the rule is accurate for hundred percent of the transactions in which bread and butter are included.
- A relationship seems to be remarkable when the relationship is recognized by the algorithm with a measure of confidence >= predefined threshold.
- This type of predefined threshold is known as ***minimum confidence***. A higher confidence denotes that the rule (X → Y) is further attractive or more reliable, depending upon the sample dataset.
- There are two common measures which are used by the Apriori algorithm: support and confidence. The ranking of all the rules can be decided depending upon these two measures for the purpose of skipping the uninteresting rules by retaining the interesting rules.
- From all the candidate rules, it is easy for confidence to sort out the interesting rules, but there is a problem;
- Given rules in the form of X → Y, confidence suppose just the ancestor (X) and the co-existence of X and Y; it does not consider the resultant of the rule (Y).
- Hence it is difficult for confidence to guess if a rule has true implication of the relationship or it is purely coincidental.
- Even though X and Y may statistically independent, their confidence score will be high. This issue is addressed by other measures like lift and leverage.
- The responsibility of Lift is to measure the count of combine occurrence of X and Y if they are independent of each other in a statistical manner.
- **Lift** is considered as a measure of the way by which X and Y are actually related instead of coincidentally coming together.

$$\text{Lift } (X \rightarrow Y) = \frac{\text{Support } (X \wedge Y)}{\text{Support } (X) \cdot \text{Support } (Y)} \quad \ldots(3.3.1)$$

- If the X and Y are statistically not dependent on each other, then value of Lift is one. If there is some usefulness to the rule then lift of X → Y is greater than 1. If the association of X and ?Y has greater strength, then the value of Lift will be larger.
- Consider 1,000 transactions in which {milk,butter} 300 times, {milk} in 500 times, and butter in 400 times. Then Lift(milk → butter) = 0.3/(0.5*0.4) = 1.5.
- If {bread} exist in 400 transactions and {milk,bread} exist in 400, then

  Lift(milk → bread) = 0.4/(0.5*0.4) = 2.
- Hence the conclusion is that the association of milk and bread is stronger as compared to the association of milk and butter.
- **Leverage** is considered as a same notion, but rather than using a ratio, leverage takes the reference of the difference.
- The difference is calculated by Leverage in the probability of X and Y existing with each other in the dataset compared to the expected situation when X and Y were statistically not dependent on each other.

Leverage (X → Y) = Support (X ^ Y) – Support (X) · Support (Y)

- Theoretically value of leverage is zero when X and Y are statistically not dependent on each other.
- The value of leverage is more than zero when X and Y have any type of relationship.

- If the value of leverages is larger, then it is considered that the relationship between X and Y is stronger.
- A larger leverage value indicates a stronger relationship between X and Y.
- For the example which we have seen previously, Leverage(milk → butter) = 0.3–(0.5*0.4) = 0.1 and Leverage(milk → bread) = 0.4–(0.5*0.4) = 0.2.
- This again proves that the association of milk and bread is stronger as compared to the association of milk and butter.
- For confidence it is possible to identify trustworthy rules, but it is unable to tell whether a rule is coincidental.
- Sometimes a high-confidence rule may be considered as ambiguous since in the rule consequent, confidence does not consider support of the itemset.
- The measures like lift and leverage assures identification of interesting rules and also filter out the coincidental rules.

Syllabus Topic : Case Study - Transactions in Grocery Store

3.4 Case Study - Transactions in Grocery Store

Q. 3.4.1 Write a case study on Transactions in Grocery Store. (Refer section 3.4) **(8 Marks)**

Objective

- Cross selling analysis is an important analytical tool in the retail industry.
- The sales of a retail store can be improved by determining the positioning of goods, and designing sales promotion plans based on product movement.
- The five month's billing data sets of a Retail Super Market were considered for the study.

Methodology

- The issues of product positioning in a well-established retail store are examined using data mining to identify the item sets that are bought frequently.
- The item sets collected from the billing database are mined using Apriori algorithm in R platform and then the association rules are generated.

Findings

- The study analyses the buying pattern of consumers in the general store and its product positioning.
- From the study it is observed that the provisionary items which are labelled with the store owned brand name are in high movement than others.

Applications/Improvement

- The outcome of the study will be useful for the retail stores for making better decisions on product positioning, increasing the sales margin and the number of store owned products or brands.

1. Introduction

- The more time and money consuming process in retail industry is extracting more useful information from point of sale database, in order to gain competitive advantage.
- India is a market zone which comprises of huge number grocery and petty shops and large number of retail stores and supermarkets.
- In this study an attempt has been made through cross selling analysis to identify the association between products and to suggest necessary product promotion strategies.
- In addition using this analysis, the customer preferences and buying pattern have been studied to suggest product positioning to improve the sales.
- The outcome of the study can be used for decisions like pricing, promotion and product assortment and to identify the frequency with which some of the unrelated products are purchased.

2. Problem Statement

- The ability of a retailer to maximize sales largely depends on the customer need should be identified by a retailer and adapted to them.

- To achieve this retailer should work on product assortment and availability of different category of products.
- By which the opportunity of the competitor can be reduced.
- But the availability of products in store should not be an Inventory burden.
- Cross selling analysis is a possible way to identify the products which can be put together and the products which are highly preferred by customers due to availability only in this store as a competitive advantage.
- Cross selling analysis provides the retailer the information about related sales of goods like sanitary napkins and baby diaper.

3. Store Description

- The real time dataset comprising of five month's daily sales was taken for analysis from a well established retail store.
- Cross selling analytical tool was used to learn more about product assortment and customer buying pattern.
- The store has seven separate sections.

(i) House hold items (ii) Fruits and vegetables
(iii) Bakery (iv) Kitchen wares
(v) Books (vi) Personal care
(vii) Staples.

- The store consists of 9055 different products and in which there is number of store owned products.
- The geographical location of the store is that it is nearer to schools, residential area and corporate offices.

4. Apriori Algorithm

- Apriori is an effective algorithm used for mining the itemsets and application of association rules on the transactional databases.
- The association rules determined by Apriori highlight the general trend in point of sale database.
- The Key problem in the retail sector is to find useful hidden patterns for business application.
- Apriori are designed to operate on database containing transactions (the collection of items by loyal customers).
- When frequent item sets are arrived at the association rules are formed with confidence larger than or equal to a user defined minimum confidence.
- Products are the entities that we are identifying relationship between. A group of products is called an item set.

$$P = \{p_1, p_2 \cdots p_n\}$$

- Transactions are instances of groups of products co-existing together.
- For a retail store, a transaction is, basically, a, transaction.
- For an online bookstore, a transaction might be the cluster of articles read in a single visit to the website.
- For each transaction, then, we have an item set.

  tn = {pi,pj,...,pk}
- Rules are statements of the form

  $p_1, p_2, \ldots\} \Rightarrow \{p_k\}$

  i.e. if you have the products in the dataset (on the left hand side (LHS) of the rule i.e. {p_1, p_2...}), then it is likely that a customer will be pursued in the products on the right hand side (RHS i.e. {p_k}.
- In our example above, our rule would be :

  Coffee powder, milk}⇒ {sugar}
- The output of a cross selling analysis is basically a set of rules, that help us to make business decisions. (E.g. Marketing, Product Assortment).

Support

- Percentage or total number of transactions that contain all of the items in an item set (e.g., coffee powder, milk and sugar).
- The rules with high support will be applicable to a large number of transactions.
- E.g., Retail supermarket is likely to involve basic products such as coffee powder, milk and sugar that are popular across entire users.

- In case of Electronics items, such as printer toner, may not have products with a high support, because each customer only buys toner for their own use and it takes log time to replace.

Confidence

- Probability that a transaction which contains the items on the LHS of the rule (coffee powder, milk) also contains the item on the RHS (sugar).
- Higher confidence value implies greater that the item on the right hand side will be purchased.

Lift

- Probability of all of the items in a rule purchasing together (also known as the support) divided by the product of the probabilities of the items on the LHS and RHS occurring as if there was no relationship between them.
- For example, if coffee powder, milk and sugar occurred together in 3.5% of all transactions, coffee powder and milk in 20% of transactions and sugar in 6% of transactions, then the lift would be :

  $0.035/(0.2*0.06) = 2.917$.

- A lift of more than 1 suggests that the presence of coffee powder and milk increases the probability that a sugar will also occur in the transaction.
- In general, lift tells about the strength of association between the products on the LHS of the rule; larger lift value implies the greater link between two products.

5. Experimental Results

Table 3.4.1 : Top 50 Items from the billing details

Output Summary	
No. of Transactions	52674
No. of Products	9055
Support (%)	0.001

- The Fig. 3.4.1 represents the top 50 items out of 9055 based on the frequency of purchase.

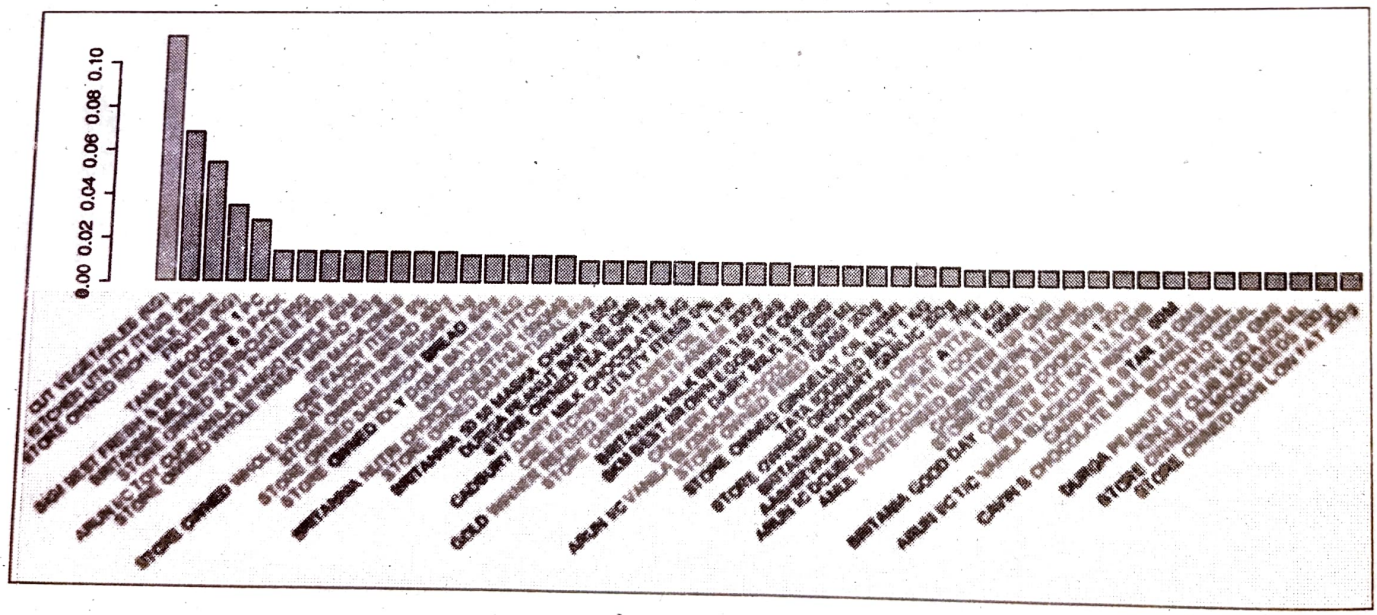

Fig. 3.4.1

- It describes the consumer buying behaviour about a particular product in that time period.
- Billing details for a period from 1st Jan. 2016 – 31st May, 2016 were considered for the cross selling analysis.
- The data set was cleaned and filtered and employed as input for the analysis process.
- Apriori algorithm is used for finding the relationship between the items and association rules generated.

Table	Confidence (%)	25
Store	Rules Generated	50

Fig. 3.4.2 : Database Output

- The given data were analyzed with different support and confidence levels.
- From the Table 1, it was identified that the rules generated were at very low support which means (0.001 * 52674 =52.674) the combination of products bought together was reflected in 53 bills.

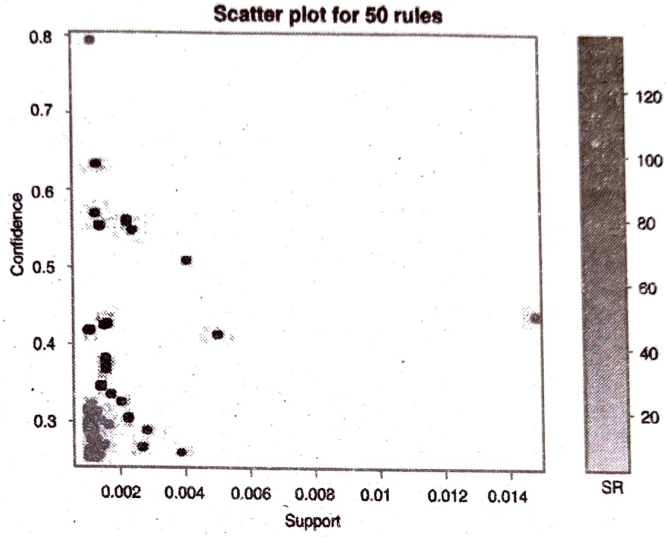

Fig. 3.4.3 : Scatter Plot

- The scatter plot displays the various products support and confidence level based on the lift value.
- This Fig. 3.4.4 represents the association between products. In it we can identify the higher number of products associated with cut vegetables.
- And there is also good number of association with store owned brands.
- So, it is suggested that the store can produce more number of variants in vegetables segment and store owned provisionary items.

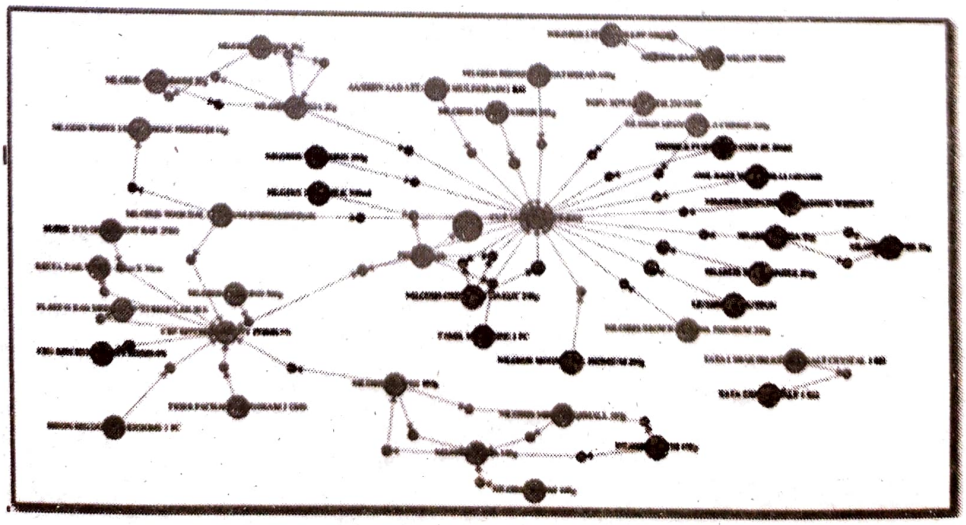

Fig. 3.4.4 : Association Chart

6. Conclusion

- Cross selling analysis using Apriori algorithm helps in identifying the frequently bought item sets and to establish the association rules for the Retail Stores.
- This analysis generally facilitates a retail store to manage the positioning of products in their store layout.
- Related products are placed along in such a way that customers can logically find the items he/she would possibly purchase such products and the same will ultimately save consumer search time and ultimately to helps to enhance turnover and the profit.
- Customers are grouped and association rules are generated separately to captivate their specific needs in a price effective manner using some special promotions.
- The results have shown that the cross selling analysis using an Apriori algorithm for retail stores improves its revenue even though the support value and number of rules are comparatively less.
- Store owned products are played a vital role in the transactions and attracts more customers and suggestion is of increasing the volume and variety of such products.

Syllabus Topic : Validation and Testing

3.5 Validation and Testing

Q. 3.5.1 Write note on Validation and Testing in Data Analytics. (Refer section 3.5) **(4 Marks)**

- It is very essential to use one or more methods for the purpose of validating the results in the business context by operating on a sample dataset.
- To establish first approach we can refer the various statistical measures like confidence, lift, and leverage.
- Rules in which there is presence of mutually independent items or the rule which are able to cover comparatively less number of transactions are considered as not interesting since they may capture false relationships.
- Confidence helps to calculate the chance when X and Y appear collectively as compared to the chance X appears.
- The interestingness of the rules can be identified by the confidence.
- The other statistical measures such as Lift and leverage are used to compare the support of X and Y opposite to their independent support.
- In the process of mining data with association rules, there may be coincidence in generation of some rules.
- E.g., if 95% of customers like to buy X while 90% of customers prefer Y, then the occurrence of X and Y jointly should be minimum 85% of the time, even though there is absence of any type of relationship between them.
- It is ensured by the lift and leverage that the identification of interesting rules will be done instead of coincidental ones.
- It is possible to establish different set of criteria with the help of subjective arguments.
- Even though the confidence is high, the rule may be assumed subjectively not interesting until it provide any unexpected profitable actions.
- E.g. rules such as {paper} → {pencil} may be subjectively considered as uninteresting or not meaningful even though there is presence of high support and confidence values.
- On the other hand, a rule like {paper}→{milk} which satisfies both of the minimum support as well as minimum confidence can be assumed as subjectively interesting since this rule is seem to be unexpected and may recommend a cross-sell chance for the retailer.
- It is difficult to incorporate subjective knowledge into the evaluation of rules, and for this purpose there is need of collaboration with domain experts.
- The role of domain experts may be as business users or the business intelligence analysts as an integral part of the team of Data Science project.

- The data Science team can interact with the results and make a decision about whether they are appropriate to operationalize them.

Syllabus Topic : Diagnostics

3.6 Diagnostics

Q. 3.6.1 Write Diagnostics of Apriori algorithm. (Refer section 3.6) **(4 Marks)**

- Even if the Apriori algorithm is considered as simple to understand as well as implement, few of the rules generated by this algorithm may be uninteresting or practically useless.
- Also generation of some rules may be because of coincidental relationships which are set between the variables.
- For the purpose of addressing this problem, one has to use measures such as confidence, lift, and leverage
- There is one more problem regarding association rules is that, in Phase 3 as well as Phase 4 of the Data Analytics Lifecycle, the team should mention the minimum support before the process of model execution, which may generate extremely more or extremely less rules.
- In the subsequent research, it is possible for a variant of the algorithm to use a default target range for the number of rules.
- It will help the algorithm to adjust the minimum support accordingly.
- Apriori algorithm is considered as one of the earliest as well as most fundamental algorithms for the purpose of generating association rules.
- The Apriori algorithm helps to lower the computational workload by the process of just examining itemsets which meet the precise minimum threshold.
- On the other hand, the Apriori algorithm may be computationally expensive which depends upon the size of the dataset.
- For each and every level of support, the algorithm needs a process of scanning of the whole database to get the result.
- The time to compute in each run increases as per increase in the database.
- We will see some approaches to enhance Apriori's efficiency :

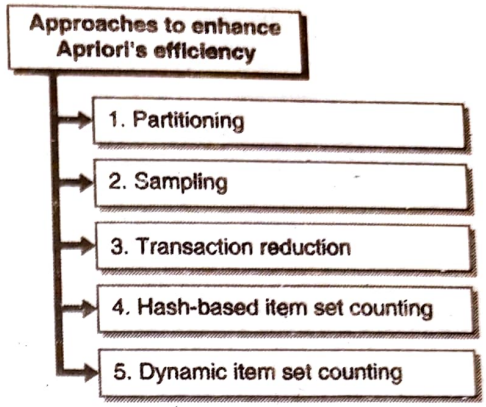

Fig. 3.6.1 : Approaches to enhance Apriori's efficiency

➜ **1. Partitioning**

- It is necessary for an item set to be frequent in minimum one partition of the transaction database if that item set is potentially frequent in a transaction database.

➜ **2. Sampling**

- It takes out a subset of the data which is having lower support threshold and takes the help of the subset to carry out association rule mining.

➜ **3. Transaction reduction**

- A transaction which does not have frequent k-itemsets is considered as useless in the process of subsequent scans and hence can be ignored.

➜ **4. Hash-based item set counting**

- If the subsequent hashing bucket count of a k-itemset is lower than a specific threshold, then it is confirm that the k-itemset cannot be frequent.

➜ **5. Dynamic itemset counting**

- Only when all of the subsets of new candidate itemsets are anticipated to be frequent, the itemsets can be added.

3.7 Regression Analysis

Q. 3.7.1 Explain Regression analysis in detail. (Refer section 3.7) **(8 Marks)**

- Lets take a simple example : Suppose your manager asked you to predict annual sales. There can be a hundred of factors (drivers) that affects sales.
- In this case, sales is your dependent variable. Factors affecting sales are independent variables. Regression analysis would help you to solve this problem.
- In simple words, regression analysis is used to model the relationship between a dependent variable and one or more independent variables.
- It helps us to answer the following questions –
  1. Which of the drivers have a significant impact on sales ?
  2. Which is the most important driver of sales ?
  3. How do the drivers interact with each other ?
  4. What would be the annual sales next year ?

☞ Terminologies related to regression analysis

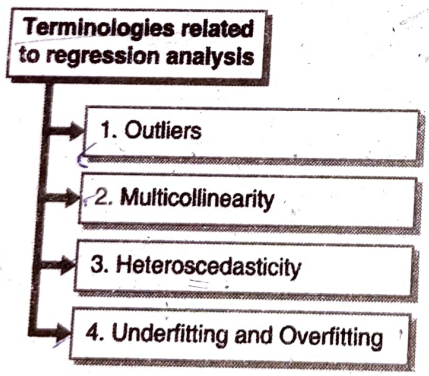

Fig. 3.7.1 : Terminologies related to regression analysis

➜ 1. Outliers

- Suppose there is an observation in the dataset which is having a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier.
- In simple words, it is extreme value. An outlier is a problem because many times it hampers the results we get.

➜ 2. Multicollinearity

- When the independent variables are highly correlated to each other then the variables are said to be multicollinear.
- Many types of regression techniques assume that multicollinearity should not be present in the dataset.
- It is because it causes problems in ranking variables based on its importance.
- Or it makes job difficult in selecting the most important independent variable (factor).

➜ 3. Heteroscedasticity

- When dependent variable's variability is not equal across values of an independent variable, it is called heteroscedasticity.
- **Example** - As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals.
- Those with higher incomes display a greater variability of food consumption.

➜ 4. Underfitting and Overfitting

- When we use unnecessary explanatory variables it might lead to overfitting.
- Overfitting means that our algorithm works well on the training set but is unable to perform better on the test sets. It is also known as problem of high variance.
- When our algorithm works so poorly that it is unable to fit even training set well then it is said to underfit the data. It is also known as problem of high bias.
- In the Fig. 3.7.2 we can see that fitting a linear regression (straight line in (1)) would underfit the data i.e. it will lead to large errors even in the training set.
- Using a polynomial fit in (2) is balanced i.e. such a fit can work on the training and test sets well, while in (3) the fit will lead to low errors in training set but it will not work well on the test set.

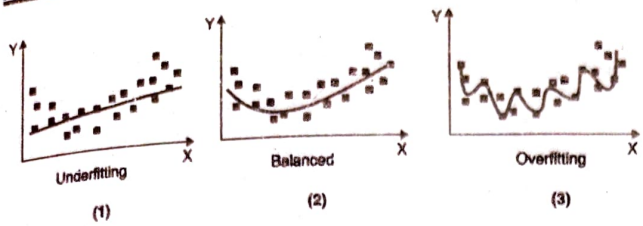

Fig. 3.7.2 : Regression – Underfitting and Overfitting

☞ Types of Regression

- Every regression technique has some assumptions attached to it which we need to meet before running analysis.
- These techniques differ in terms of type of dependent and independent variables and distribution.

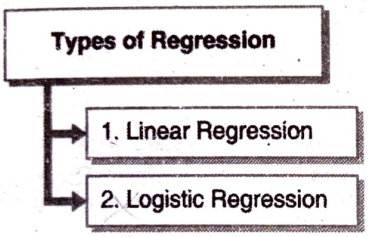

Fig. 3.7.3 : Types of Regression

Syllabus Topic : Linear Regression

3.7.1 Linear Regression

Q. 3.7.2 Explain Linear Regression. (Refer section 3.7.1) **(8 Marks)**

- It is the simplest form of regression. It is a technique in which the dependent variable is continuous in nature.
- The relationship between the dependent variable and independent variables is assumed to be linear in nature.
- We can observe that the given plot represents a somehow linear relationship between the mileage and displacement of cars.
- The points are the actual observations while the line fitted is the line of regression.
- When you have only 1 independent variable and 1 dependent variable, it is called simple linear regression.
- When you have more than 1 independent variable and 1 dependent variable, it is called Multiple linear regression.

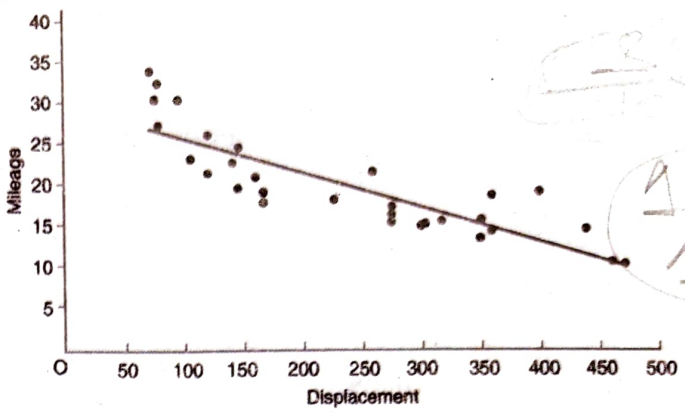

Fig. 3.7.4 : Regression Analysis

- The equation of multiple linear regression is listed below :

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

- **Multiple Regression Equation :** Here 'y' is the dependent variable to be estimated, and X are the independent variables and ε is the error term. βi's are the regression coefficients.

☞ Assumptions of linear regression

1. There must be a linear relation between independent and dependent variables.
2. There should not be any outliers present.
3. No heteroscedasticity.
4. Sample observations should be independent.
5. Error terms should be normally distributed with mean 0 and constant variance.
6. Absence of multicollinearity and auto-correlation.

☞ Estimating the parameters

- To estimate the regression coefficients βi's we use principle of least squares which is to minimize the sum of squares due to the error terms i.e.

$$\text{Min}\Sigma\varepsilon^2 = \text{Min}\Sigma\,[y - (\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k)]^2$$

- On solving the above equation mathematically we obtain the regression coefficients as :

$$\hat{\beta} = (X'X)^{-1}X'y$$

- **Interpretation of regression coefficients :** Let us consider an example where the dependent variable is marks obtained by a student and explanatory variables

are number of hours studied and number of classes attended. Suppose on fitting linear regression we got the linear regression as :

- Marks obtained = 5 + 2 (no. of hours studied) + 0.5(no. of classes attended)
- Thus we can have the regression coefficients 2 and 0.5 which can interpreted as :
  1. If number of hours studied and number of classes are 0 then the student will obtain 5 marks.
  2. Keeping number of classes attended constant, if student studies for one hour more then he will score 2 more marks in the examination.
  3. Similarly keeping number of hours studied constant, if student attends one more class then he will attain 0.5 marks more.

☞ Linear Regression in R

- We consider the swiss data set for carrying out linear regression in R. We use lm() function in the base package. We try to estimate Fertility with the help of other variables.

```
library(datasets)
model = lm(Fertility ~ .,data = swiss)
lm_coeff = model$coefficients
lm_coeff
summary(model)
```

- The output we get is :

```
> lm_coeff
(Intercept)  Agriculture  Examination  Education  Catholic
66.9151817 -0.1721140  -0.2580082  -0.8709401  0.1041153
Infant.Mortality
  1.0770481
> summary(model)
Call:
lm(formula = Fertility ~ ., data = swiss)
Residuals:
   Min     1Q  Median     3Q    Max
-15.2743 -5.2617  0.5032  4.1198 15.3213
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.91518  10.70604  6.250 1.91e-07 ***
Agriculture  -0.17211   0.07030 -2.448  0.01873 *
Examination  -0.25801   0.25388 -1.016  0.31546
Education    -0.87094   0.18303 -4.758 2.43e-05 ***
Catholic      0.10412   0.03526  2.953  0.00519 **
Infant.Mortality 1.07705  0.38172  2.822  0.00734 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

- Hence we can see that 70% of the variation in Fertility rate can be explained via linear regression.

Syllabus Topic : Logistic Regression

3.7.2 Logistic Regression

Q. 3.7.3 Explain Logistic Regression. **(8 Marks)**
(Refer section 3.7.2)

- In logistic regression, the dependent variable is binary in nature (having two categories). Independent variables can be continuous or binary. In multinomial logistic regression, you can have more than two categories in your dependent variable.
- Why don't we use linear regression in some case? The homoscedasticity assumption is violated.
- Errors are not normally distributed
- y follows binomial distribution and hence is not normal.

☞ Examples

☞ HR Analytics

- IT firms recruit large number of people, but one of the problems they encounter is after accepting the job offer many candidates do not join.

- So, this results in cost over-runs because they have to repeat the entire process again.
- Now when you get an application, can you actually predict whether that applicant is likely to join the organization (Binary Outcome - Join / Not Join).

Elections

- Suppose that we are interested in the factors that influence whether a political candidate wins an election.
- The outcome (response) variable is binary (0/1); win or lose.
- The predictor variables of interest are the amount of money spent on the campaign and the amount of time spent campaigning negatively.
- Predicting the category of dependent variable for a given vector X of independent variables. Through logistic regression we have

$$P(Y=1) = \exp(a + B \square X) \ / (1 + \exp(a + B \square X))$$

- Thus we choose a cut-off of probability say 'p' and if $P(Yi = 1) > p$ then we can say that Yi belongs to class 1 otherwise 0.

Interpreting the logistic regression coefficients (Concept of Odds Ratio)

- If we take exponential of coefficients, then we'll get odds ratio for ith explanatory variable.
- Suppose odds ratio is equal to two, then the odds of event is 2 times greater than the odds of non-event.
- Suppose dependent variable is customer attrition (whether customer will close relationship with the company) and independent variable is citizenship status (National / Expat).
- The odds of expat attrite is 3 times greater than the odds of a national attrite.

Logistic Regression in R

- In this case, we are trying to estimate whether a person will have cancer depending whether he smokes or not.
- We fit logistic regression with **glm( )** function and we set **family = "binomial".**

```
model <- glm(Lung.Cancer..Y.~Smoking..X.,data = data,
family = "binomial")
```

- The predicted probabilities are given by :

#Predicted Probablities

model$fitted.values

```
        1         2         3         4         5         6         7         8         9
0.4545455 0.4545455 0.6428571 0.6428571 0.4545455 0.4545455 0.4545455 0.4545455 0.6428571
       10        11        12        13        14        15        16        17        18
0.6428571 0.4545455 0.4545455 0.6428571 0.6428571 0.6428571 0.4545455 0.6428571 0.6428571
       19        20        21        22        23        24        25
0.6428571 0.4545455 0.6428571 0.6428571 0.4545455 0.6428571 0.6428571
```

- Predicting whether the person will have cancer or not when we choose the cut off probability to be 0.5

```
data$prediction <- model$fitted.values>0.5
> data$prediction
 [1] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
[16] FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
```

Syllabus Topic : Reasons to Choose and Cautions

3.7.3 Reasons to Choose and Cautions

Q. 3.7.4 What are reasons to choose a particular regression and cautions? (Refer section 3.7.3) **(8 Marks)**

- Linear regression is considered appropriate in the situation when the input variables are in the form of continuous or discrete, with categorical data types, however the outcome variable is continuous.
- Logistic regression is suitable when the outcome variable is categorical.
- In both the modules, linear additive function of the input variables is taken into account.
- The performance of both regression techniques will be poor in case such an assumption does not hold true
- Additionally, in linear regression, the supposition of usually distributed error terms with a constant variance is considered as significant for number of the statistical inferences which can be taken into account.
- If it is found that the several assumptions do not appear to hold, it is necessary to apply appropriate transformations on the data.
- Even if a set of input variables is a good forecaster for the outcome variable, the analyst should not suppose that the input variables straightforwardly result an outcome.
- E.g., it may be recognized that the people who have usual dentist visits may have a less threat of heart attacks.
- Though, just sending an individual to the dentist almost surely has no impact on the person's possibility of having a heart attack.
- It is likely that frequent dentist visits may point to a person's overall health as well as dietary choices, which may show direct effect on a person's health.
- This example describes the usually known expression, "Correlation does not imply causation."
- One has to take caution while applying previously fitted model to data which falls outside the dataset used to train the model.
- In a regression model, it may be possible that linear relationship may no longer hold at values outside the training dataset.
- E.g., consider as income an input variable and the values of income are in between Rs.50,000 to Rs.100,000, In such case applying the model to incomes which are not in the range could result in incorrect estimates and predictions.
- In a linear regression model, the state of residence gives an easy additive term to the income model; however there will not be any effect on the coefficients of the other input variables, like Age and Education.
- On the other hand, if state makes impact on the other variables' effect to the income model, substitute approach will generate fifty distinct linear regression models: one model for each state.
- This approach is considered as an example of the options as well as decisions which the data scientist must be eager to consider.
- Multicollinearity is a condition when lot of input variables highly associated to each other.
- Multicollinearity may usually generate coefficient estimates which are comparatively big in absolute magnitude and may be of wrong direction (-ve or +ve sign).
- Whenever possible, it is necessary to remove majority of these associated variables from the model or replaced by a new variable.
- E.g., in a regression application related to medical field, height and weight may are important input variables, but it may be possible that these variables tend to be correlated.
- In such case, one can use the BMI (Body Mass Index), which is a function of height and weight of an individual.
- BMI = weight / height where weight is in kilograms and height is in meters.

Syllabus Topic : Additional Regression Models

3.7.4 Additional Regression Models

Q. 3.7.5 Explain types of Regression Analysis. (Refer section 3.7.4) (8 Marks)

- Except linear and logistic regression models, there are some more regression models :

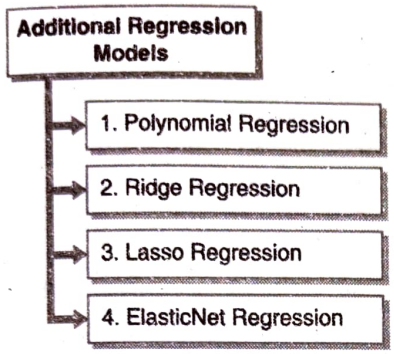

Fig. 3.7.5 : Additional Regression Models

➜ 1. Polynomial Regression

- When we want to create a model that is suitable for handling non-linearly separable data, we will need to use a polynomial regression.
- In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points. For a polynomial regression, the power of some independent variables is more than 1.
- For example, we can have something like:

$$Y = a\_1*X\_1 + (a\_2)^2*X\_2 + (a\_3)^4*X\_3 \ldots\ldots a\_n*X\_n + b$$

☞ A few key points about Polynomial Regression

- Able to model non-linearly separable data; linear regression can't do this. It is much more flexible in general and can model some fairly complex relationships.
- Full control over the modelling of feature variables (which exponent to set).
- Requires careful design. Need some knowledge of the data in order to select the best exponents.
- Prone to over fitting if exponents are poorly selected.

➜ 2. Ridge Regression

- A standard linear or polynomial regression will fail in the case where there is high collinearity among the feature variables.
- Collinearity is the existence of near-linear relationships among the independent variables. The presence of high collinearity can be determined in a few different ways:
- A regression coefficient is not significant even though, theoretically, that variable should be highly correlated with Y.
- When you add or delete an X feature variable, the regression coefficients change dramatically.
- Your X feature variables have high pairwise correlations.
- We can first look at the optimization function of a standard linear regression to gain some insight as to how ridge regression can help :

$$\min \| Xw - y \|^2$$

- where X represents the feature variables, w represents the weights, and y represents the ground truth. Ridge Regression is a remedial measure taken to alleviate collinearity amongst regression predictor variables in a model.
- Collinearity is a phenomenon in which one feature variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.
- Since the feature variables are so correlated in this way, the final regression model is quite restricted and rigid in its approximation i.e it has high variance.
- To alleviate this issue, Ridge Regression adds a small squared bias factor to the variables :

$$\min \| Xw—y \|^2 + z\| w \|^2$$

- Such a squared bias factor pulls the feature variable coefficients away from this rigidness, introducing a small amount of bias into the model but greatly reducing the variance.

A few key points about Ridge Regression

- The assumptions of this regression are same as least squared regression except normality is not to be assumed.
- It shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection.

3. Lasso Regression

- Lasso Regression is quite similar to Ridge Regression in that both techniques have the same premise.
- We are again adding a biasing term to the regression optimization function in order to reduce the effect of collinearity and thus the model variance.
- However, instead of using a squared bias like ridge regression, lasso instead uses an absolute value bias:

$$\min \| Xw - y \|^2 + z\| w \|$$

4. ElasticNet Regression

- ElasticNet is a hybrid of Lasso and Ridge Regression techniques. It uses both the L1 and L2 regularization taking on the effects of both techniques:

$$\min \| Xw.- y \|^2 + z\_1\| w \| + z\_2\| w \|^2$$

- A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation.

A few key points about ElasticNet Regression

- It encourages group effect in the case of highly correlated variables, rather than zeroing some of them out like Lasso.
- There are no limitations on the number of selected variables.

❑❑❑