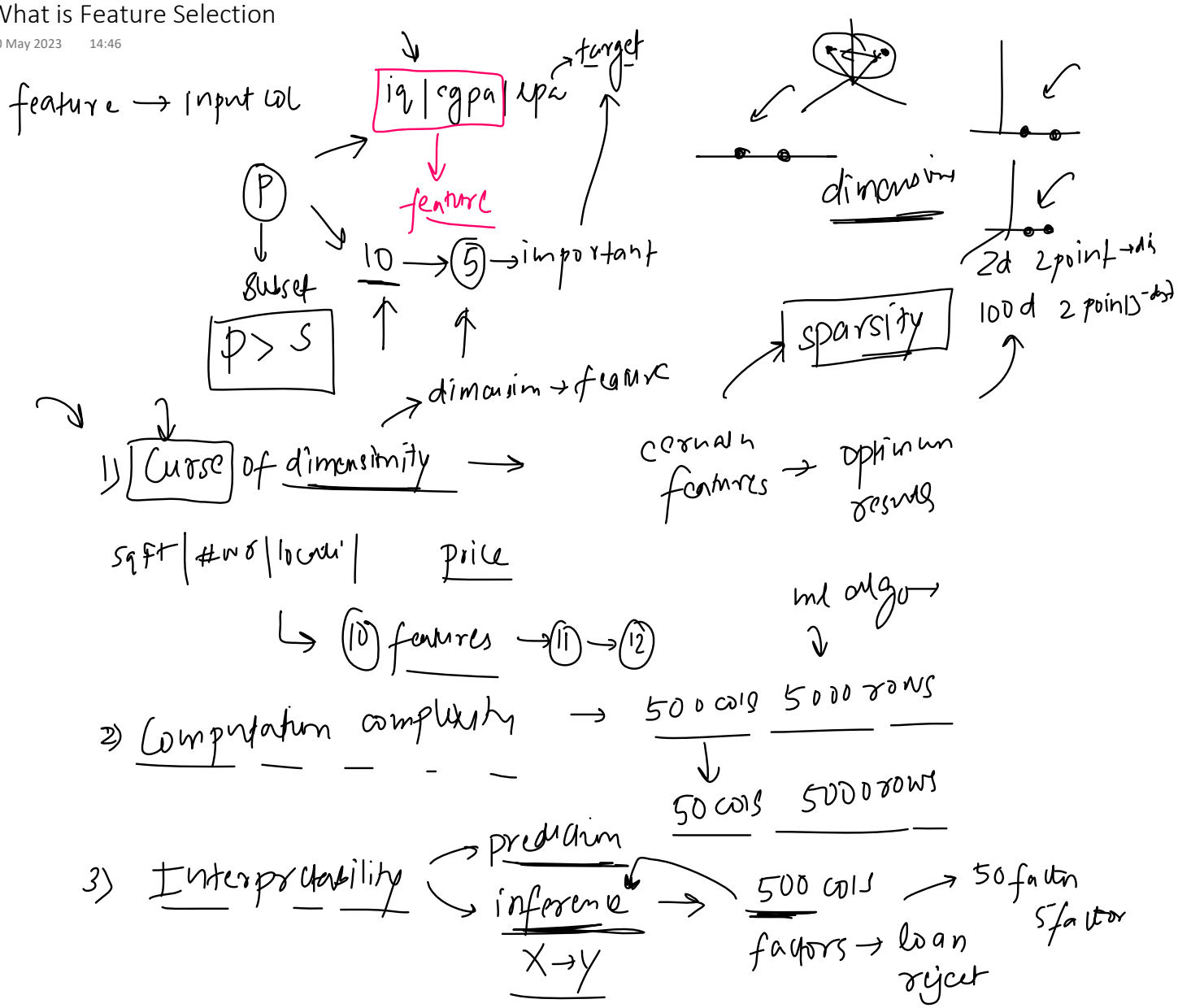
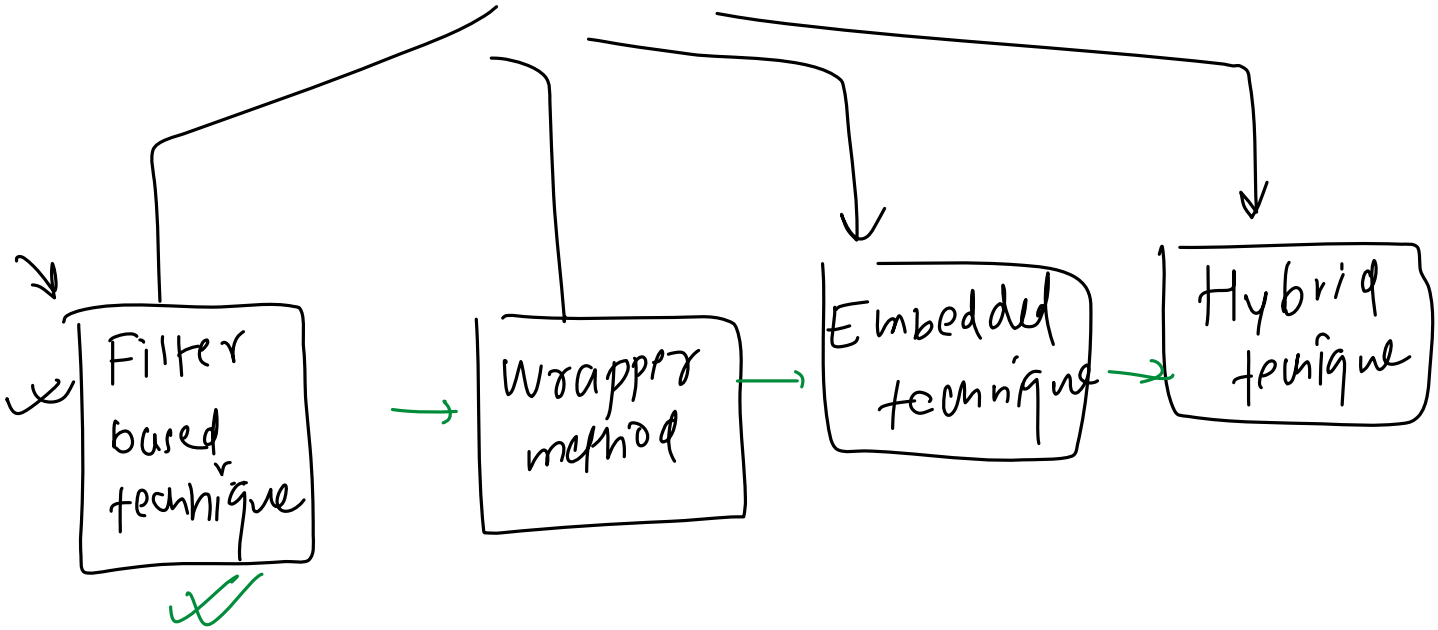
What is Feature Selection

10 May 2023 14:46



Types of Feature Selection

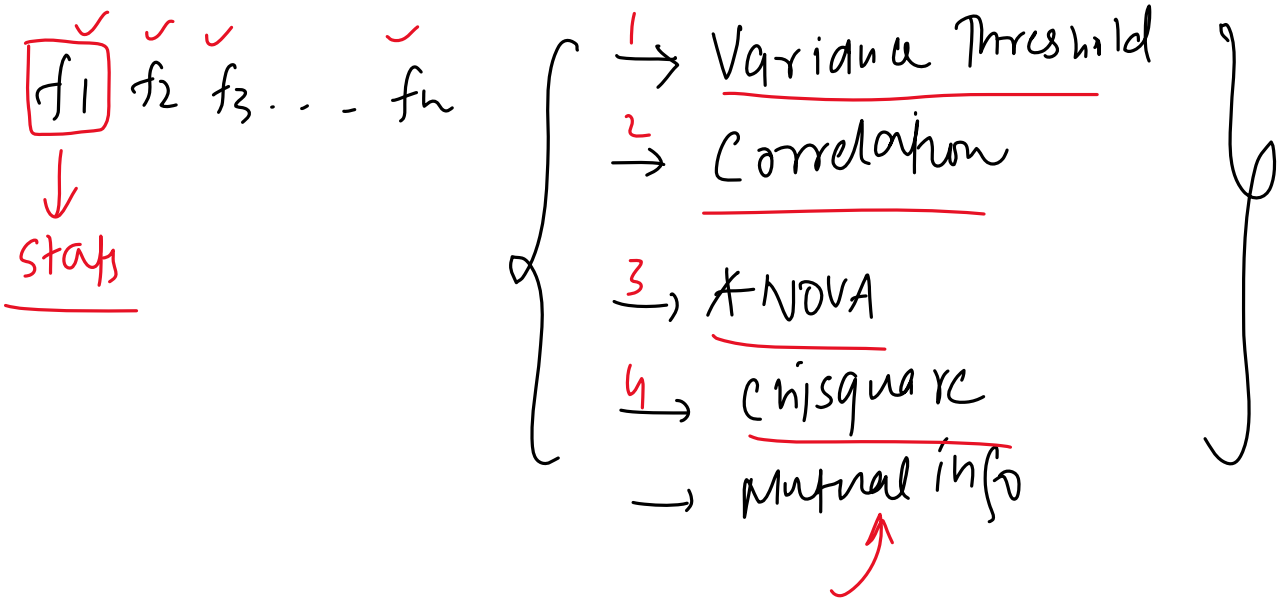
10 May 2023 14:47



Filter Based Feature Selection

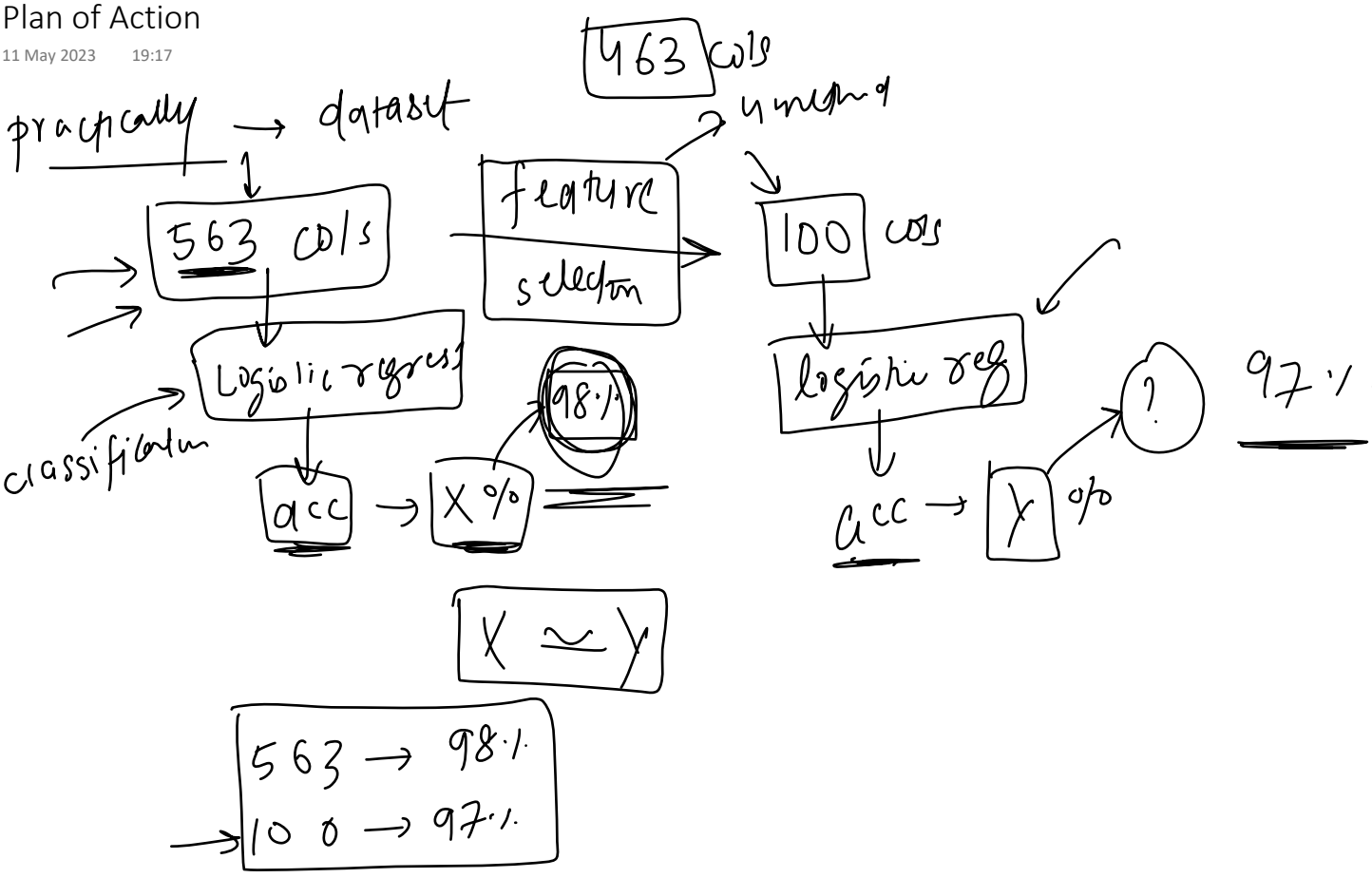
10 May 2023 14:47

Filter-based feature selection techniques are methods that use statistical measures to score each feature independently, and then select a subset of features based on these scores. These methods are called "filter" methods because they essentially filter out the features that do not meet some criterion.



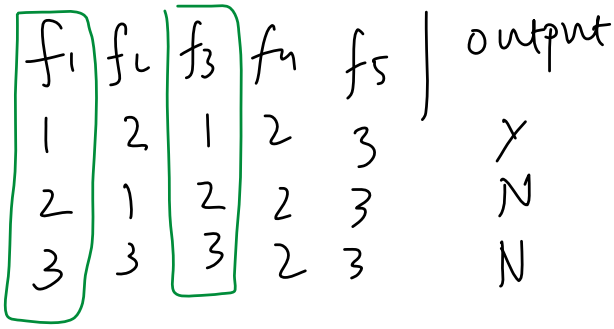
Plan of Action

11 May 2023 19:17



1. Duplicate Features

10 May 2023 14:47



same

dict → keep keys → dict → values

key → original cols

values → duplicate

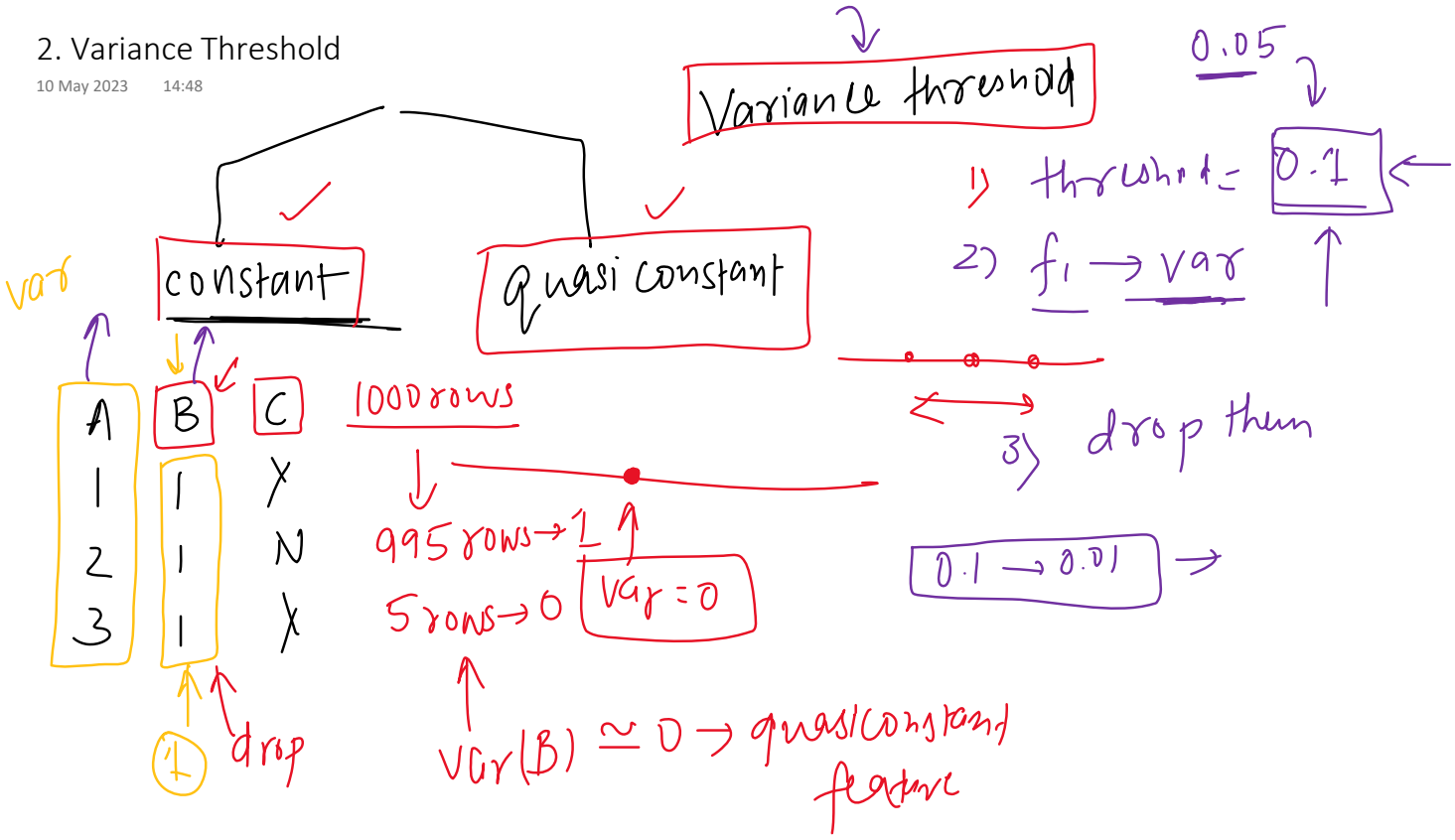
561 → 540

↓

21 duplicate

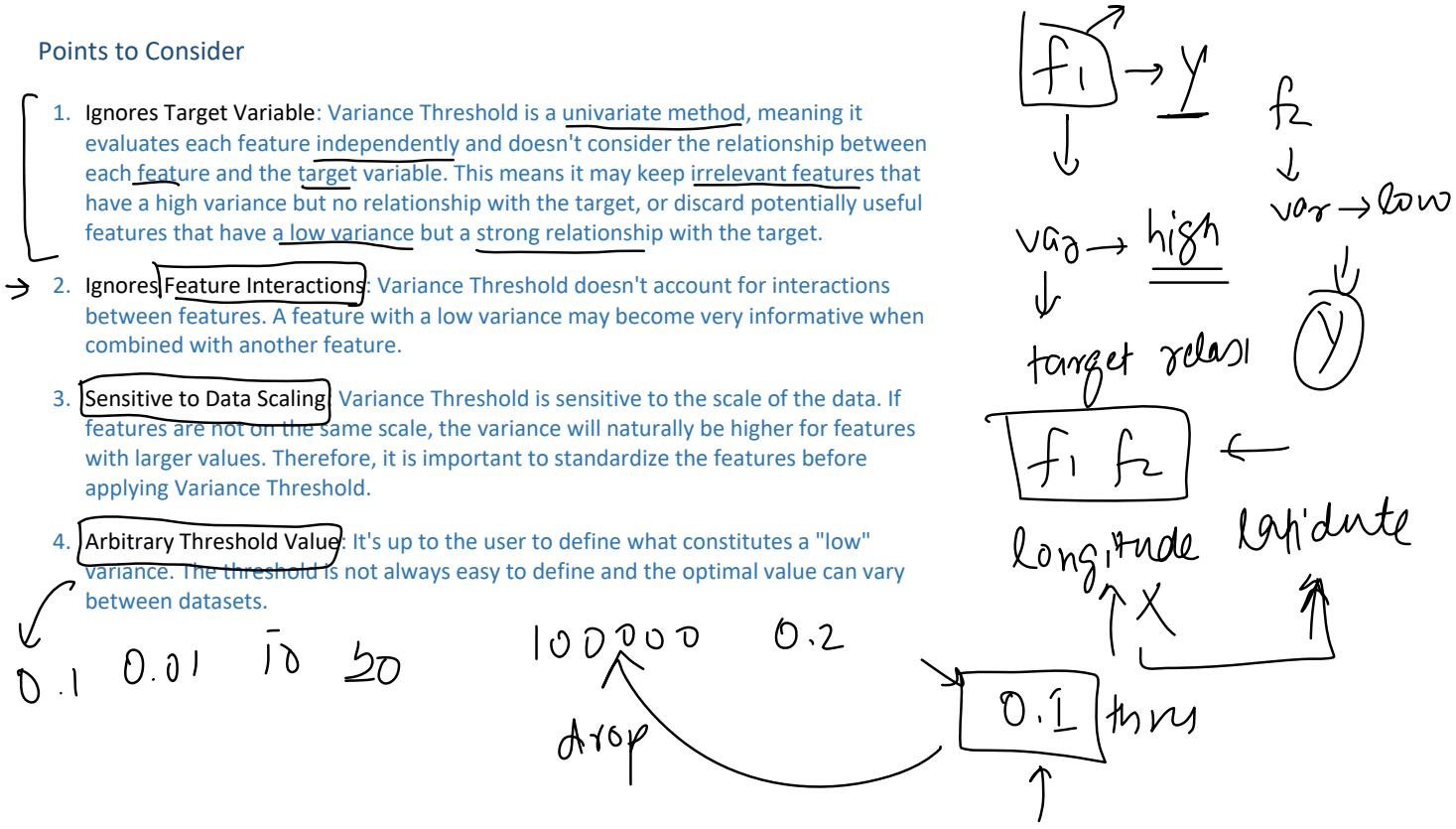
2. Variance Threshold

10 May 2023 14:48



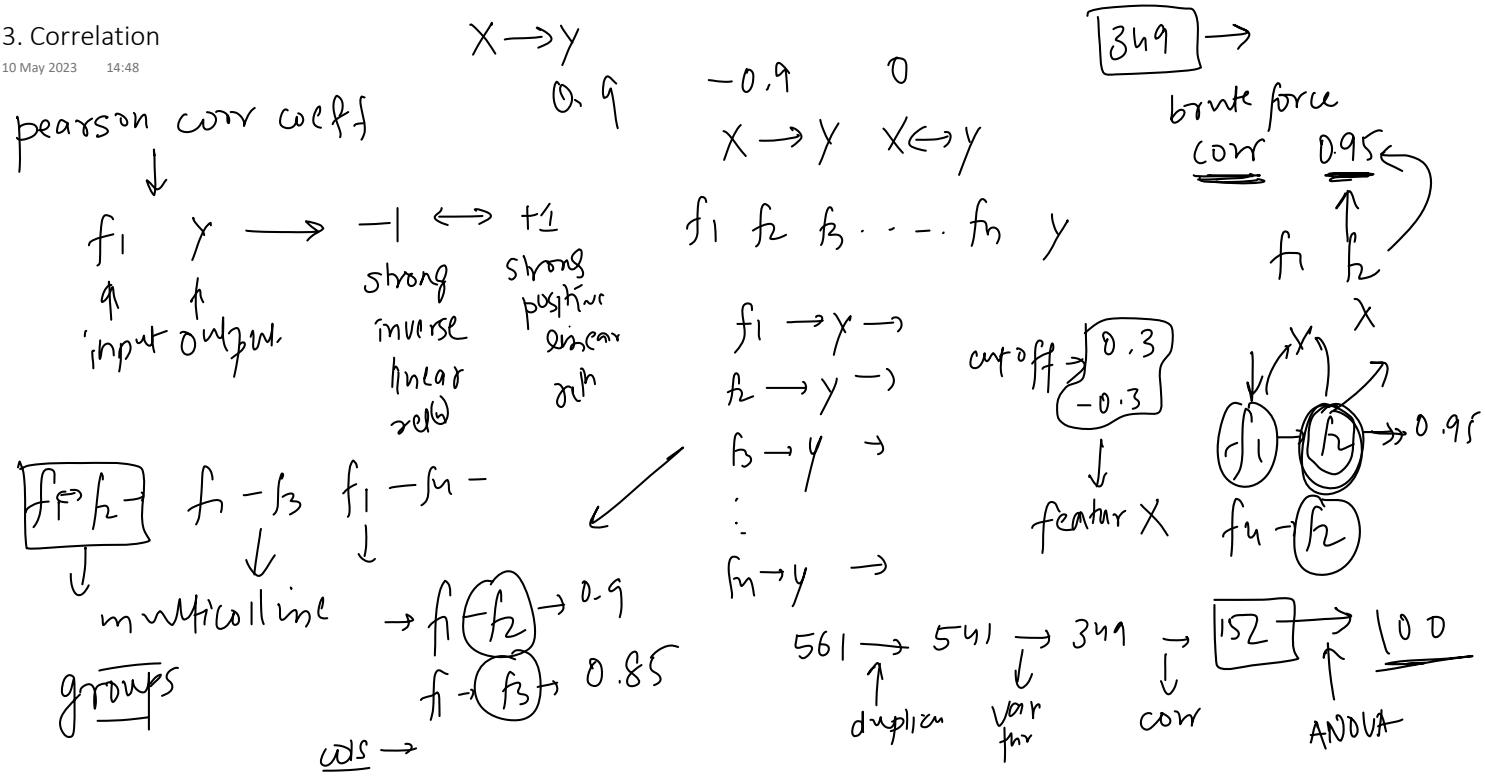
Points to Consider

1. Ignores Target Variable: Variance Threshold is a univariate method, meaning it evaluates each feature independently and doesn't consider the relationship between each feature and the target variable. This means it may keep irrelevant features that have a high variance but no relationship with the target, or discard potentially useful features that have a low variance but a strong relationship with the target.
2. Ignores Feature Interactions: Variance Threshold doesn't account for interactions between features. A feature with a low variance may become very informative when combined with another feature.
3. Sensitive to Data Scaling: Variance Threshold is sensitive to the scale of the data. If features are not on the same scale, the variance will naturally be higher for features with larger values. Therefore, it is important to standardize the features before applying Variance Threshold.
4. Arbitrary Threshold Value: It's up to the user to define what constitutes a "low" variance. The threshold is not always easy to define and the optimal value can vary between datasets.



3. Correlation

10 May 2023 14:48



Disadvantages

1. Linearity Assumption: Correlation measures the linear relationship between two variables. It does not capture non-linear relationships well. If a relationship is nonlinear, the correlation coefficient can be misleading.
2. Doesn't Capture Complex Relationships: Correlation only measures the relationship between two variables at a time. It may not capture complex relationships involving more than two variables.
3. Threshold Determination: Just like variance threshold, defining what level of correlation is considered "high" can be subjective and may vary depending on the specific problem or dataset.
4. Sensitive to Outliers: Correlation is sensitive to outliers. A few extreme values can significantly skew the correlation coefficient.

4. ANOVA

10 May 2023 14:49

f1 → y rel? ↑ strong

k-1 n-k

n → # rows k → # categories (6)

→ hypothesis test → numerical f1 → Y categorical → >2 classes

1 way ANOVA

numerical f1 → y numerical

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio
Between samples or categories	$n_1(\bar{X}_1-\bar{\bar{X}})^2+\cdots+n_k(\bar{X}_k-\bar{\bar{X}})^2$	$(k-1)$	MS between: SS between / $(k-1)$	MS between / MS within
Within samples or categories	$\sum(X_{1i}-\bar{X}_1)^2+\cdots+\sum(X_{ki}-\bar{X}_k)^2$ $i=1,2,3,\ldots$	$(n-k)$	MS within: SS within / $(n-k)$	
Total	$\sum(X_{ij}-\bar{\bar{X}})^2$ $i,j=1,2,3,\ldots$	$(n-1)$		

f-statistic → p-value → num f → y

f1 f2 . . . fn52 → Y (cat > 2) (num)

ANOVA

→ f1 → y → no rel? → skl num

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS) (This is SS Divided by d.f.) and is an Estimation of Variance to be Used in F-ratio	F-ratio
Between samples or categories	$n_1(\bar{X}_1-\bar{\bar{X}})^2+\cdots+n_k(\bar{X}_k-\bar{\bar{X}})^2$	$(k-1)$	SS between / $(k-1)$	MS between / MS within
Within samples or categories	$\sum(X_{1i}-\bar{X}_1)^2+\cdots+\sum(X_{ki}-\bar{X}_k)^2$ $i=1,2,3,\ldots$	$(n-k)$	SS within / $(n-k)$	
Total	$\sum(X_{ij}-\bar{\bar{X}})^2$ $i,j=1,2,3,\ldots$	$(n-1)$		

SSW = 50

$(1-X_S)^2+(4-X_S)^2+(3-X_S)^2+(2-X_L)+(5-X_L)^2+(6-X_L)^2+(3-X_W)^2+(7-X_W)^2+(8-X_W)^2$

f1 → y f1 +acc X

1 s
2 l
3 w
4 s

S: 1, 4, 3 → X_S

L: 2, 5, 6 → X_L

W: 3, 7, 8 → X_W

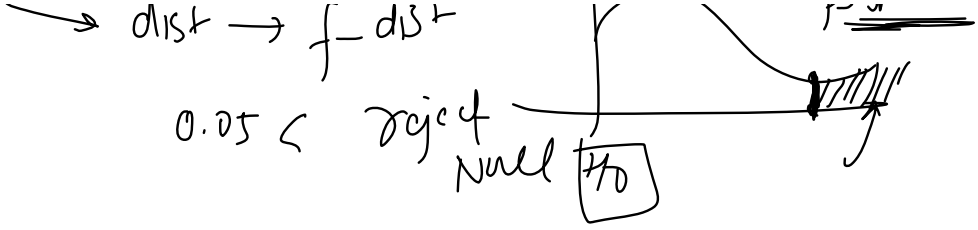
grand mean $\bar{\bar{X}}$

SSW / K-1 / SSB / n-k = f-ratio

var

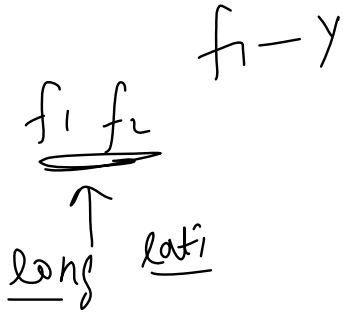
→ $SSB = 3(X_S-\bar{\bar{X}})^2+3(X_L-\bar{\bar{X}})^2+3(X_W-\bar{\bar{X}})^2$

dist → f-dist → p-value



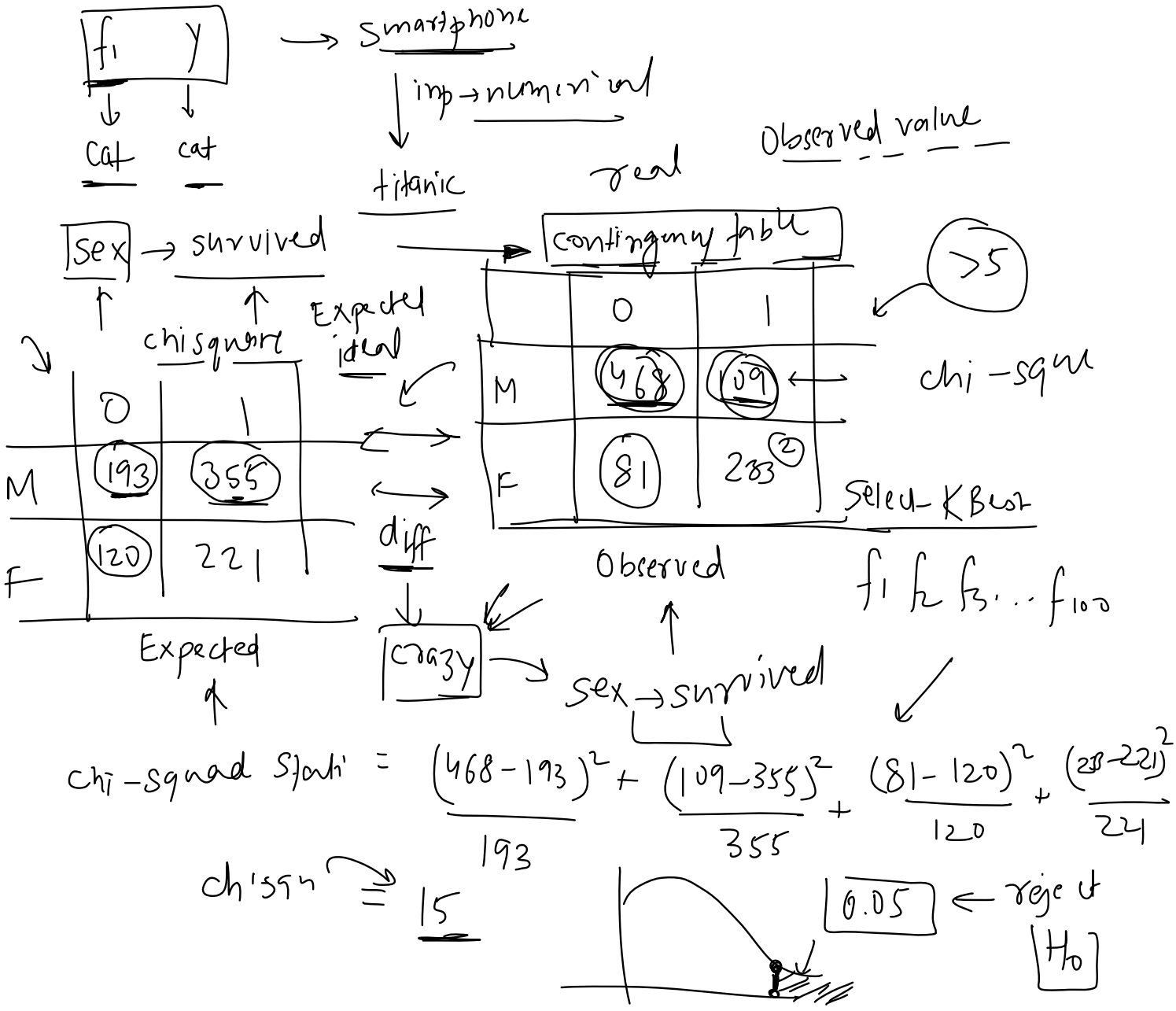
Disadvantages

1. Assumption of Normality: ANOVA assumes that the data for each group follow a normal distribution. This assumption may not hold true for all datasets, especially those with skewed distributions.
2. Assumption of Homogeneity of Variance: ANOVA assumes that the variances of the different groups are equal. This is the assumption of homogeneity of variance (also known as homoscedasticity). If this assumption is violated, it may lead to incorrect results.
3. Independence of Observations: ANOVA assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
4. Effect of Outliers: ANOVA is sensitive to outliers. A single outlier can significantly affect the F-statistic leading to a potentially erroneous conclusion.
5. Doesn't Account for Interactions: Just like other univariate feature selection methods, ANOVA does not consider interactions between features.



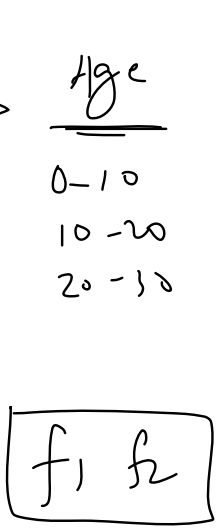
5. Chi-Square

10 May 2023 14:48



Disadvantages

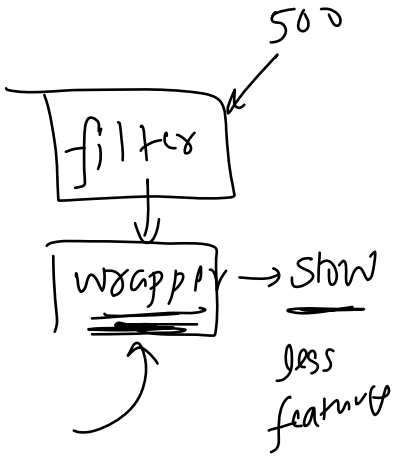
1. Categorical Data Only: The chi-square test can only be used with categorical variables. It is not suitable for continuous variables unless they have been discretized into categories, which can lead to loss of information.
2. Independence of Observations: The chi-square test assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (e.g., time series data, nested data).
3. Sufficient Sample Size: Chi-square test requires a sufficiently large sample size. The results may not be reliable if the sample size is too small or if the frequency count in any category is too low (typically less than 5).
4. No Variable Interactions: Chi-square test, like other univariate feature selection methods, does not consider interactions between features. It might miss out on identifying important features that are significant in combination with other features.



$f_1 \rightarrow y$

Advantages and Disadvantages

11 May 2023 16:07



Advantages

1. Simplicity: Filter methods are generally straightforward and easy to understand. They involve calculating a statistic that measures the relevance of each feature, and selecting the top features based on this statistic.
2. Speed: These methods are usually computationally efficient. Because they evaluate each feature independently, they can be much faster than wrapper methods or embedded methods, which need to train a model to evaluate feature importance.
3. Scalability: Filter methods can handle a large number of features effectively because they don't involve any learning methods. This makes them suitable for high-dimensional datasets.
4. Pre-processing Step: They can serve as a pre-processing step for other feature selection methods. For instance, you could use a filter method to remove irrelevant features before applying a more computationally expensive method, such as a wrapper method.

Disadvantages

1. Lack of Feature Interaction: Filter methods treat each feature individually and hence do not consider the interactions between features. They might miss out on identifying important features that don't appear significant individually but are significant in combination with other features.
2. Model Agnostic: Filter methods are agnostic to the machine learning model that will be used for the prediction. This means that the selected features might not necessarily contribute to the accuracy of the specific model you want to use.
3. Statistical Measures Limitation: The statistical measures used in these methods have their own limitations. For example, correlation is a measure of linear relationship and might not capture non-linear relationships effectively. Similarly, variance-based methods might keep features with high variance but low predictive power.
4. Threshold Determination: For some methods, determining the threshold to select features can be a bit subjective. For example, what constitutes "low" variance or "high" correlation might differ depending on the context or the specific dataset.