**Preprocessing for Raw Data**

1. Deleting the irrelevant variables
2. Filling in the missing data

|  |  |
| --- | --- |
| **Demographics** | |
| count | Remove due to the irrelevant information |
| **gender** | 1. There is no significant difference on gender in chi-square test 2. Random fill:   probability for “male” = 0.506; probability for “female” = 1-0.506 |
| **age**, under 30, senior citizen | 1. Fill in “under 30” and “senior citizen” by “age” 2. Fill with mean = 24 in “age” for those are under 30 3. Fill with mean = 72 in “age” for those are senior citizen 4. Fill with total mean = 47 in “age” for missing all data |
| **married** | 1. There is no significant difference on gender in chi-square test 2. Random fill in “married”:   probability for “No” = 0.518; probability for “Yes” = 1-0.518 |
| dependents,  **number of dependents** | 1. Fill in “dependents” by “number of dependents” 2. Fill with the highest frequency “2” for only missing data in “number of dependents” 3. Fill with the highest frequency “0” in “number of dependents” for missing data both in “dependents” and “number of dependents” since there is no significant difference among “number of dependents” |
| **Location** | |
| country, state, zip code, | Remove because all the zip codes are clustered in southern California in the San Diego area |
| city | Remove due to the similar information with la long |
| **lat long**, **latitude, longitude** | 1. Fill in “lat long” by “latitude” and “longitude”, and vice versa 2. Fill in “lat long” by “zip code” and “ID” |
| **Population** | |
| **ID, zip code** | Remove because all the zip codes are clustered in southern California in the San Diego area |
| **population** | Fill in “population” by “ID” |
| **Sample Submission** | |
| **churn category** | Remove those are missing data in training set |
| **Satisfaction** | |
| **satisfaction score** | Fill in by forward and backward method |
| **Services** | |
| count | Remove due to the irrelevant information |
| quarter | Remove because all the quarters are the same |
| referred a friend,  **number of referrals** | 1. Fill in “referred a friend” by “number of referrals” 2. Fill with the highest frequency “1” in “number of referrals” for those have referred a friend 3. Fill with the highest frequency “0” in “number of referrals” for missing data both in “referred a friend” and “number of referrals” since there is no significant difference on “number of referrals” under each churn category |
| **phone service**,  **multiple lines** | 1. Fill with “Yes” in “phone service” for those have “None” or “Fiber Optic” in “internet type” 2. Fill in the remaining “phone service” by random sampling of binomial distribution with probability = 0.904 3. Fill with “No” in “multiple lines” for those have “No” in “phone service” 4. Fill in the remaining “multiple lines” by random sampling of multinomial distribution of following probabilities   Probability = 0.566 for “Bank Withdrawal”  Probability = 0.874 for “Credit Card”  Others for “Mailed Check” |
| **tenure in months**, **offer** | 1. Fill in “tenure in months”, “total long-distance charges”, “total charges”, “avg long-distance charges”, and “monthly charges” by these functions.   Tenure in months = total long-distance charges / avg long-distance charges  Tenure in months = total charges / monthly charges   1. Fill in the remaining “tenure in months” by the mean of each “offer”   Mean(A)= 70; Mean(B) = 53; Mean(C) = 31; Mean(D) = 16; Mean€ = 4   1. Fill in the remaining “tenure in months” by random sampling in the beta distribution for “None” and remaining missing data in “offer”   Beta distribution: alpha = 0.526; beta = 0.673   1. Fill with “None” in “offer” for the remaining missing data |
| **internet service,**  **online security,**  **online backup,**  **device protection plan, premium tech support, streaming TV,**  **streaming movies, streaming music** | 1. Fill with “Yes” in “internet service” for those have “No” in “phone service” 2. Fill with “Yes” in “internet service” for those have “Yes” in “multiple lines”, “avg long-distance charges”, and “total long-distance charges” 3. Fill with “No” in “internet service”, “online security”, “online backup”, “device protection plan”, “premium tech support”, “streaming TV”, “streaming movies”, “streaming music”, “unlimited data” for those have “None” in “internet type” 4. Fill with “Yes” in “internet service” for those have “internet type” or have “avg monthly GB download” larger than zero 5. Fill with “No” in “internet service” for those have “No” in all “online security”, “online backup”, “device protection plan”, “premium tech support”, “streaming TV”, “streaming movies”, “streaming music”, “unlimited data” 6. Fill with “Yes” in “internet service” for those have at least one “Yes” in “online security”, “online backup”, “device protection plan”, “premium tech support”, “streaming TV”, “streaming movies”, “streaming music”, “unlimited data” 7. Fill in “internet service” by random sampling of binomial distribution with probability of 0.777   Fill in the remaining “online security”, “online backup”, “device protection plan”, “premium tech support”, “streaming TV”, “streaming movies”, and “streaming music” by forward and backward method |
| **internet type** | 1. Fill with “fiber optic” when “monthly charge” >90 2. Fill with “None” when “monthly charge” <25 3. Fill with “DSL” in the remaining data by random sampling of binomial distribution with probability of 0.655; otherwise, fill with “Cable”” |
| **avg monthly GB download, unlimited data** | 1. Fill with “0” in “avg monthly GB download” for those have “None” in “internet type” 2. Fill in “avg monthly GB download” by forward and backward method |
| **contract, paperless billing, payment method** | 1. Fill in “contract” by random sampling of multinomial distribution with the following probabilities   probability = 0.505 for “Month-to-Month”  probability = 0.45 for “One Year”  others for “Two Year”   1. Fill with ‘Yes’ in “paperless billing” by random sampling of multinomial distribution with the probability of 0.59 2. Fill in “payment method” by random sampling of multinomial distribution with the following probabilities   probability = 0.566 for “Bank Withdrawal”  probability = 0.874 for “Credit Card”  others for “Mailed Check” |
| **Satisfaction score** | 1. Fill in by forward and backward method |
| **total long-distance charges,**  **total charges,**  **avg long-distance charges, monthly charges**  **total refunds, total extra data charges, total revenue** | 1. Fill with “0” in “avg monthly long-distance” for those have “No” in “phone service” 2. Fill in “total long-distance charges”, “total charges”, “avg long-distance charges”, and “monthly charges” by these functions.   Tenure in months = total long-distance charges / avg long-distance charges  Tenure in months = total charges / monthly charges   1. Fill in “total long-distance charges”, “total charges”, “total refunds”, “total extra”, and “total revenue” by the following function:   Total revenue  = total charges + total long-distance charges + total extra – total refunds   1. Fill with “0” in the remaining “total refunds”, and modified the negative values of “total extra” as “0” when it was less than 0.05. |
| **Status** | |
| **Churn Category** | Remove those are missing data in training set |