II Dynamic ordering:

|  |  |  |
| --- | --- | --- |
| Option | Algorithm | Pro & Con |
| Method 1  (correlation based) | **Step1:** Subset the whole cohort’s summary frequency table based on the known features and demographic information from user’s input.  **Step 2:** Calculate correlation scores (Phi coefficient) between each pair of unknown features by using the crosstab frequencies.  **Step 3:** Select the feature having the highest correlations (max sum of scores) with the rest, which is the first or next feature to be asked.  **Step 4:** Rank this feature’s correlation scores with others in descending order, which is used as the order for the rest of features if necessary\*.  *\*The next feature set could be just one or multiple which will depend on the design of API.* | To find the feature most correlated to the rest among those unknown features upon data resulted from dynamic sub-setting  **Pros:**  Typical inferential analysis to look into statistical significant  Specific sampling based on user’s input  Explores relationship between two or multiple features based on the rate of occurrence of a given events.  Reduces running into multicollinearity since each time the most influenced feature is removed from the question list.  **Cons:**  Does not provide directional hypothesis – meaning there is no explanation about what features drive the prediction. Thus the order may not be helpful for model prediction.  The “third” variable problem occurs – there might be other outbound influences to the correlation between each pair or multiple pairs of the defined features.  Tends to bring in bias if asking related features in certain unclear sequence |
| Method 2  (correlation based) | **Step1:** Subset the whole cohort’s summary frequency table based on the known features and demographic information from user’s input.  **Step 2, 3:**  **Condition I:** If all features are unknown,  Follow the steps 2 & 3 in Method 1  **Condition II:** If at least one feature was known, then to calculate correlation scores between each of the left features and all the known features by using the crosstab frequencies.  Select the feature having the highest correlations (max sum of scores) with all known features, which is the next feature to ask.  **Step 4:**  **Option I:** Follow the step 4 in Method 1  **Option II:** Rank the rest of unknown feature’s correlation scores with all known ones in decreasing order, which is used as the order for the rest of features if necessary\*. | To find the feature most correlated to the known features in a dynamic sub-setting  **Pros:**  Typical inferential analysis to look into statistical significant  Specific sampling based on user’s input  Explores relationship between two or multiple features based on the rate of occurrence of a given events.  Helps to smooth the dialog flow since the next question would be a highly related one (vs. those have been asked).  **Cons:**  Does not provide directional hypothesis – meaning there is no explanation about what features drive the prediction. Thus the order may not be helpful for model prediction.  The “third” variable problem occurs – there might be other outbound influences to the correlation between each pair or multiple pairs of the defined features.  Tends to bring in bias since more related features will be asked in sequential order. |
| Method 3  (correlation based) | A simplified version of method 2.  Instead of calculating correlations to “all” known features, only calculating correlation to the “previous” one asked.  Keep comparison and ranking steps the same. | It’s less practical than method 2 when multiple features would be grouped and asked in a same time window.  Pros and Cons: the same as above. |
| Method 4 & 5  (correlation based) | A reverse version of method 1 & 2  Instead of selecting the feature having the “highest” summary of correlation scores, selecting the feature having the “lowest” summary scores. | **Pros:**  Typical inferential analysis to look into statistical significant  Specific sampling based on user’s input  Explores relationship between two or multiple features based on the rate of occurrence of a given events.  May help to gain unknown information faster – since the most un-related one might be the most needed one. Helps to impute the missing features more precisely in earlier stage, so that to speed up prediction convergence.  **Cons:**  Does not provide directional hypothesis – meaning there is no explanation about what features drive the prediction. Thus the order may not be helpful for model prediction.  The “third” variable problem occurs – there might be other outbound influences to the correlation between each pair or multiple pairs of the defined features.  Breaks the natural dialog flow by jumping to and from unrelated topics. |
| Method 7  (correlation based) | Instead of computing correlation matrix among features, computing correlation between a specific outcome (Los, revision, complication or post recovery) with the unknown features.  Comparing and ranking in descending order. | To find the feature most related to the outcome  **Pros:**  Typical inferential analysis to look into statistical significant  Specific sampling based on user’s input  Explores relationship between two or multiple features based on the rate of occurrence of a given events.  Partially helps with getting better model prediction because the covariance between predictor and outcome is involved.  **Cons:**  Does not provide directional hypothesis – meaning there is no explanation about what features drive the prediction. Thus the order may not be helpful for model prediction.  The “third” variable problem occurs – there might be other outbound influences to the correlation between each pair or multiple pairs of the defined features.  It may not appear to be natural doc-patient dialog/ clinical operational workflow. |
| Method 8  (frequency based) | A copy version of method 1, 2 & 7, but use the frequency counts instead of correlation. | To find the feature having the most/least occurrences in a given event by comparing the other features regardless of having or not having the outcome.  **Can add extra condition** - to find the feature having the most/least occurrences in a given event by comparing the other features when having the specific outcome.  **Pros:**  Typical descriptive analysis to look into statistical significant  Helps with the imputation process because it’s a similar conception for conducting the clustering and finding the suitable clusters (imputing algorithm).  If there is outcome involved, helps with the model prediction.  **Cons:**  May not be representative when event is sparse in certain subset  Summary counts are specific to Explorys data only  More isolated single feature pickup without checking the relationship with others – no covariance info.  Breaks the natural dialog workflow, no connections from one context to another, confuses users in experience. |
| Method 9  (clinical significant based) | Acquire an order based on the clinical/surgical practice/knowledge.  **Step 1:** select the first or next feature according to the static reference being given.  **Step 2:** rank the rest unknown features based on the static order being given.  **Step 3:** maintain the static reference/order being given. | **Pros:**  The order reflects the relationships among the feature diagnosis and healthcare workflow  It would be the best one to accommodate the doctor-patient dialog from patient experience perspective.  **Cons:**  Not clear definition and standard in current  Lack of connection with the data, then the statistical model  May not be helpful to improve the predictive performance since the current model is based on a specific population, which may not be representative in general. |

Method 1 & 2:

Dynamically select a subset of the cohort based on the user’s inputs at each of the questioning step, applied with all known features;

Based on the subset, calculate the correlations between each pair among all unknown features;

Select the feature having the max/min summary of correlation value with the rest, return as the next feature to ask

Rank the other unknown features based on their correlation value with the next feature, return as the next feature list to ask

If the subset size runs under 30 patients, then use the Z value order to rank the unknown features, return the top one as the next feature

If there are more than one unknown features having the same max/min summary of correlation value, then use the Z value order to rank them, return the top one as the next feature

Method 3 & 4:

Dynamically select a subset of the cohort based on the user’s inputs at each of the questioning step, applied with all known features;

Based on the subset, calculate the frequency of each unknown features;

Select the feature having the max/min frequency count, return as the next feature to ask

Rank the other unknown features based on their frequency count, return as the next feature list to ask

In extreme cases described as above, applied with the same method by ordering with the Z value of the model

Method 5 & 6:

Dynamically select a subset of the cohort based on the user’s inputs at each of the questioning step, applied with the feature previously asked only;

Repeat Method 3 & 4

Method 7 & 8:

Dynamically select a subset of the cohort based on the user’s inputs at each of the questioning step, applied with the feature previously asked only;

Repeat Method 1 & 2

Method 9 & 10:

Dynamically select a subset of the cohort based on the user’s inputs at each of the questioning step, applied with the feature previously asked only;

If all features are unknown or there is only one known features, then

Based on the subset, calculate the frequency of each unknown features;

Select the feature having the max/min frequency count, return as the first feature to ask

Else,

Based on the subset, calculate the correlations between each pair of one unknown features and one known features which previously asked, except the one asked at the very last step, which is used to define the dynamic subset

Select the feature having the max/min summary of correlation value with the rest, return as the next feature to ask

Rank the other unknown features based on their summary of correlation value with the known features

Method 11 & 12

Dynamically select a subset of the cohort based on the user’s inputs at each of the questioning step, applied with the feature previously asked only;

Based on the subset, calculate the correlations between each pair among all features except the one currently applied to extract the subset;

Select the feature having the max/min summary of correlation value with the rest, return as the next feature to ask

Rank the other unknown features based on their correlation value with the next feature, return as the next feature list to ask

|  |  |  |
| --- | --- | --- |
| Method | Step | prediction | predicted range | error | error ratio | Comparison |
| Static  (z value based) |  |  |
| Methods 1 & 2  (correlation based, dynamic subset applied with all known features) |  |  |
| Methods 7 & 8  (correlation based, dynamic subset applied with a single feature currently answered) |  |  |
| Methods 3 & 4  (frequency based, dynamic subset applied with all known features) |  |  |
| Methods 5 & 6  (frequency based, dynamic subset applied with a single feature currently answered) |  |  |
| Methods 9 & 10  (correlation based between unknown and known, dynamic subset applied with a single feature currently answered) |  |  |
| Method 11 & 12 |  |  |