Machine Learning

Assignment 4

Team Members

AXB210119	Abhinava Bharamasagara Nanjundaiah
HXD220007	Harsha Priya Daggubati
PXP210104	Pritika Priyadarshini

Our Approach

We are given a Markov Network involving evidence, query and hidden variables along with functions involving the variables. Our dataset provides assignments for evidence and query variables. We divide the dataset into train and test. Using the training data assignments for evidence and query variables, we generate the most probable assignments for query variables.

In order to get the assignment for query variable with just evidence variable given, the accuracy may not be good enough, so the task is to increase the features in the given Markov network. The given dataset has some missing data/hidden variables whose values are estimated using Variable Elimination method and the generated values are the new features on which we train our model.

Flow of Algorithm

1. Generating Hidden assignments using bucket elimination which is described in detail below.
2. Cleaning Query Variables such as if it contains single class ex: Sample_4_2022 has 3 queries with single class.
3. Training Models with Hidden Assignments.
    a. Random Forest
    b. Logistic Regression
4. Training Models without Hidden Assignments.
    a. Random Forest
    b. Logistic Regression
5. Computing Error of Each model
6. Computing Score between relevant models.
    a. Score of LR **\`without\`** hidden vs Random Forest **\`without\`** Hidden Assignments
    b. Score of LR **\`with\`** hidden vs Random Forest **\`without\`** Hidden Assignments
    c. Score of LR **\`with\`** hidden vs Random Forest **\`with\`** hidden Assignments.

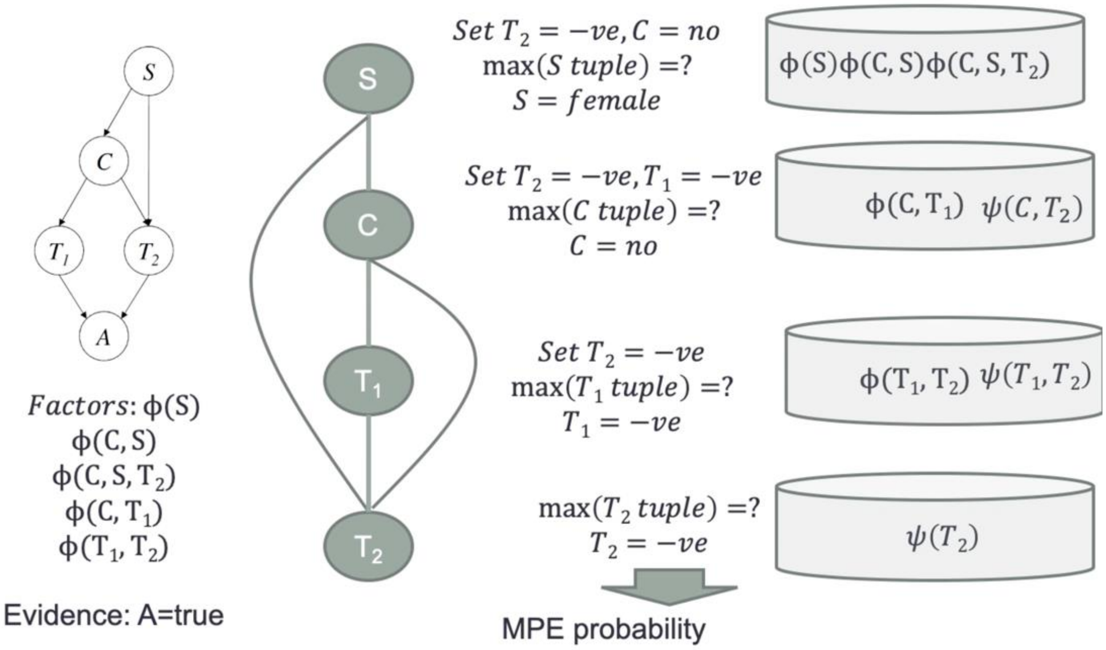
Variable Elimination

Using bucket elimination for getting assignments to hidden variables:

1. First, we run the algorithm along min-fill ordering and create buckets based on the order and put the functions into appropriate buckets.
2. After downward pass is complete, we then run the algorithm bottom up to find optimal assignments to the hidden variables. At each bucket we find an assignment to the bucket variable that has the maximum probability given an optimal assignment to all variables ordered below the current bucket, using the following steps:

   a. At the current bucket, we have a function which is defined over some variable (say T2) and we find the value of this variable that has the maximum probability in its CPT.
   b. As we go up along the order and reach the next bucket (say next variable is T1), we set the assignments for the known hidden variables (T2 in this case) and instantiate the functions in the bucket, to get a function containing only T1 so we can repeat step b.
   c. We follow steps a and b, until all the hidden variables are assigned a value.

3. After the above two steps are complete, we can augment the dataset by assigning values to the hidden variables by using the Markov network.
4. We use the new dataset obtained, with assignments to hidden variables to train a classifier (Multi Output Logistic Regression Classifier).



Testing the model

We compare the performance of Markov model with the results from a trivial solver such as random forest classifier. Err is the error of the Markov model on the test set and MaxErr is the error for the trivial solver. We compute the score using the following formula -

$$Score = \max\left(0, 100\left(1 - \frac{Err}{MaxErr}\right)\right)$$

Specifically, we compare

   a. Score of LR `**without**` hidden vs Random Forest `**without**` Hidden Assignments
   b. Score of LR `**with**` hidden vs Random Forest `**without**` Hidden Assignments
   c. Score of LR `**with**` hidden vs Random Forest `**with**` hidden Assignments

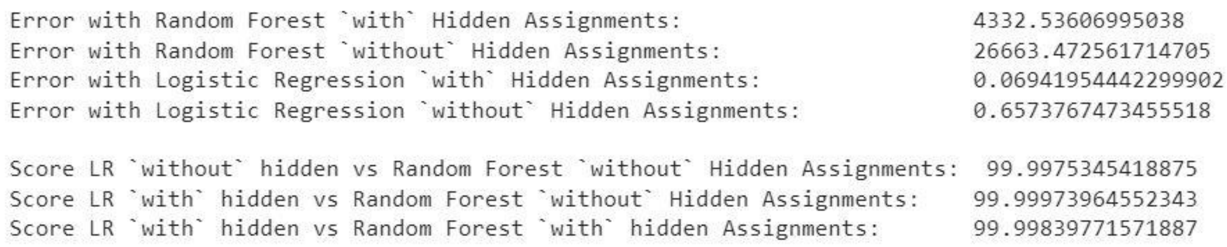
Results

We ran by using the entire test set which has roughly 3300 samples.

Dataset: Sample_1_MLC_2022

Error	Without Hidden Variables	With Hidden Variables
Random Forest	26663.472	4332.536
Logistic Solver	0.6574	0.0694

Score	LR with hidden variables	LR without hidden variables
Random Forest without hidden variables	99.9997	99.9975
Random Forest with hidden variables	99.9984	**N/A**



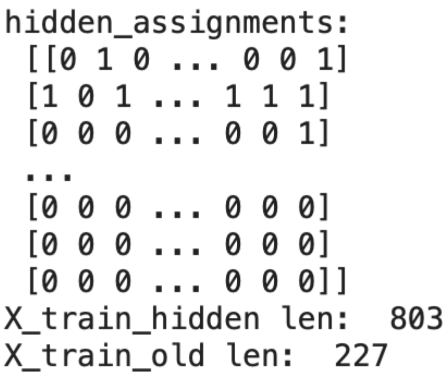
```
Error with Random Forest `with` Hidden Assignments:                     4332.53606995038
Error with Random Forest `without` Hidden Assignments:                  26663.472561714705
Error with Logistic Regression `with` Hidden Assignments:               0.06941954442299902
Error with Logistic Regression `without` Hidden Assignments:            0.6573767473455518

Score LR `without` hidden vs Random Forest `without` Hidden Assignments:  99.9975345418875
Score LR `with` hidden vs Random Forest `without` Hidden Assignments:    99.99973964552343
Score LR `with` hidden vs Random Forest `with` hidden Assignments:       99.99839771571887
```

Dataset: Sample_2_MLC_2022

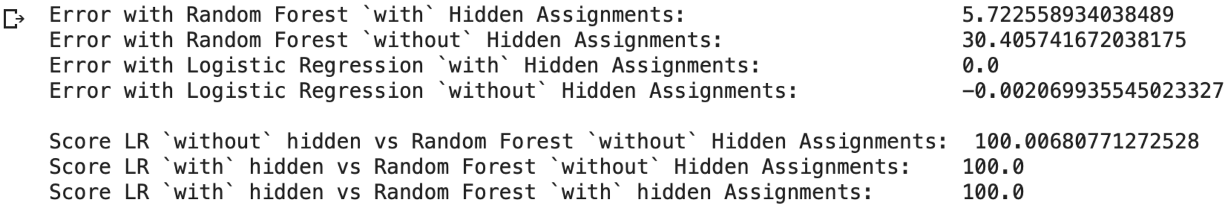
For 10 test points:

Error	Without Hidden Variables	With Hidden Variables
Random Forest	30.4057	5.7226
Logistic Solver	-0.0021	0.0

Score	LR with hidden variables	LR without hidden variables
Random Forest without hidden variables	100	100.0068
Random Forest with hidden variables	100	**N/A**



```
hidden_assignments:
 [[0 1 0 ... 0 0 1]
 [1 0 1 ... 1 1 1]
 [0 0 0 ... 0 0 1]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
X_train_hidden len:  803
X_train_old len:  227
```

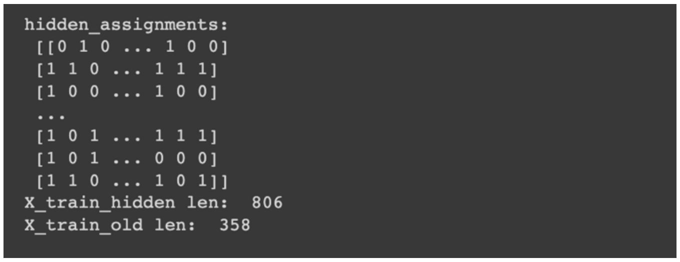


```
Error with Random Forest `with` Hidden Assignments:                     5.722558934038489
Error with Random Forest `without` Hidden Assignments:                  30.405741672038175
Error with Logistic Regression `with` Hidden Assignments:               0.0
Error with Logistic Regression `without` Hidden Assignments:            -0.002069935545023327

Score LR `without` hidden vs Random Forest `without` Hidden Assignments:  100.00680771272528
Score LR `with` hidden vs Random Forest `without` Hidden Assignments:    100.0
Score LR `with` hidden vs Random Forest `with` hidden Assignments:      100.0
```

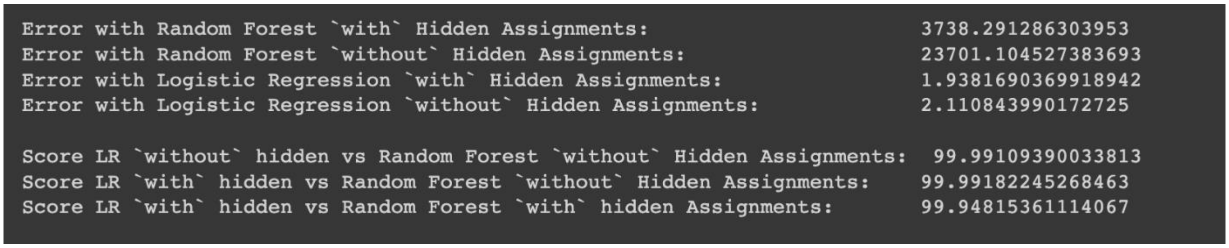
Dataset: Sample_3_MLC_2022

Error	Without Hidden Variables	With Hidden Variables
Random Forest	23701.104	3738.291
Logistic Solver	2.110	1.938

Score	LR with hidden variables	LR without hidden variables
Random Forest without hidden variables	99.9918	99.9910
Random Forest with hidden variables	99.9482	**N/A**



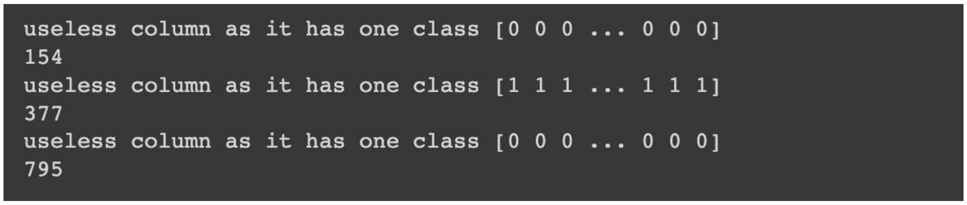
```
hidden_assignments:
 [[0 1 0 ... 1 0 0]
 [1 1 0 ... 1 1 1]
 [1 0 0 ... 1 0 0]
 ...
 [1 0 1 ... 1 1 1]
 [1 0 1 ... 0 0 0]
 [1 1 0 ... 1 0 1]]
X_train_hidden len:  806
X_train_old len:  358
```



```
Error with Random Forest `with` Hidden Assignments:                     3738.291286303953
Error with Random Forest `without` Hidden Assignments:                  23701.104527383693
Error with Logistic Regression `with` Hidden Assignments:               1.9381690369918942
Error with Logistic Regression `without` Hidden Assignments:            2.110843990172725

Score LR `without` hidden vs Random Forest `without` Hidden Assignments:  99.99109390033813
Score LR `with` hidden vs Random Forest `without` Hidden Assignments:    99.99182245268463
Score LR `with` hidden vs Random Forest `with` hidden Assignments:      99.94815361114067
```

Dataset: Sample_4_MLC_2022

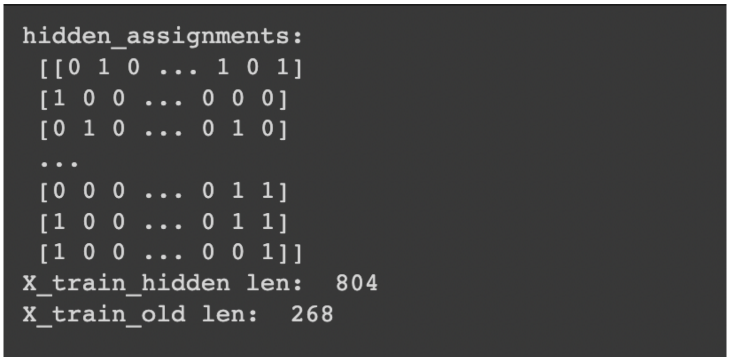
In this dataset, there were 3 query variables whose assignments were a single class (either all 0s or all 1s). Since these were meaningless assignments, we removed them from the set of query variables before training our model.



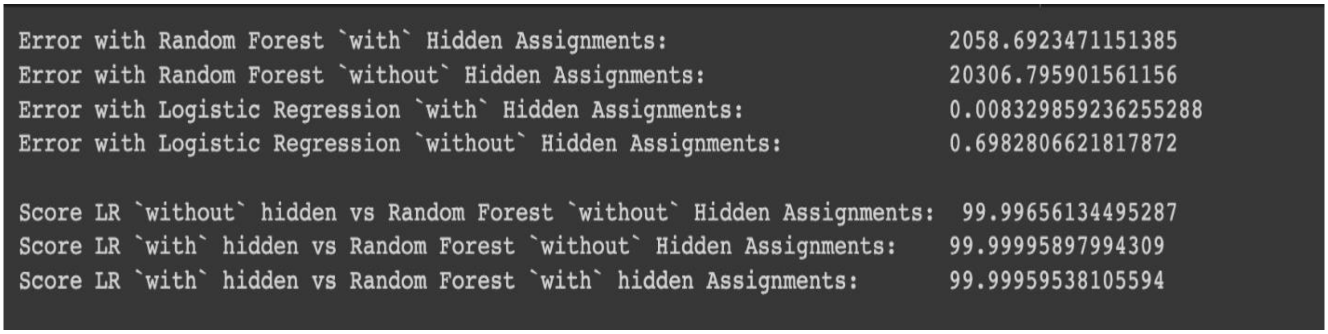
```
useless column as it has one class [0 0 0 ... 0 0 0]
154
useless column as it has one class [1 1 1 ... 1 1 1]
377
useless column as it has one class [0 0 0 ... 0 0 0]
795
```

Error	Without Hidden Variables	With Hidden Variables
Random Forest	20306.796	2058.692
Logistic Solver	0.698	0.008

Score	LR with hidden variables	LR without hidden variables
Random Forest without hidden variables	99.9999	99.9966
Random Forest with hidden variables	99.9996	**N/A**



```
hidden_assignments:
 [[0 1 0 ... 1 0 1]
 [1 0 0 ... 0 0 0]
 [0 1 0 ... 0 1 0]
 ...
 [0 0 0 ... 0 1 1]
 [1 0 0 ... 0 1 1]
 [1 0 0 ... 0 0 1]]
X_train_hidden len:  804
X_train_old len:  268
```



```
Error with Random Forest `with` Hidden Assignments:                 2058.6923471151385
Error with Random Forest `without` Hidden Assignments:              20306.795901561156
Error with Logistic Regression `with` Hidden Assignments:           0.008329859236255288
Error with Logistic Regression `without` Hidden Assignments:        0.6982806621817872

Score LR `without` hidden vs Random Forest `without` Hidden Assignments:  99.99656134495287
Score LR `with` hidden vs Random Forest `without` Hidden Assignments:     99.99995897994309
Score LR `with` hidden vs Random Forest `with` hidden Assignments:        99.99959538105594
```

Observations

1. Sample 4 contains some one class query variables which were ignored while computing.
2. Sample 2 contains less evidence variables (227) compared to other samples. Likelihood of overfitting was high in this case, so we tried reducing max_iter and changing C to compensate for the same but the result had little effect.
3. Sample 4 showed the least error rate.
4. Apart from sample 2 all the Dataset had good impact with our new features.

Conclusion

1. With our new features we can see improvements in Score in all the 4 Samples.
2. Error of Random Forest drastically reduces when our new features or hidden assignments are used.