Learning the Differences Between Data Scientists

Anirban Chowdhury
Advised by Prof. Stephanie Rosenthal and Reid Simmons

INTRODUCTION

We sought to develop insights into the processes that novice and expert data scientists use to solve problems.We presented subjects with a data science problem and an API with modeling capabilities. We used information in the lines of code they wrote in order to determine what steps they took. We then used Markov Decision Processes to capture the actions used and analyzed the learned probabilities. We also used the features we engineered from the data do develop a machine learning pipeline to predict actions. Our final goal is to develop an agent to provide recommendations to data scientists who are stuck during a stage of the problem solving process.

METHODS

Parsing

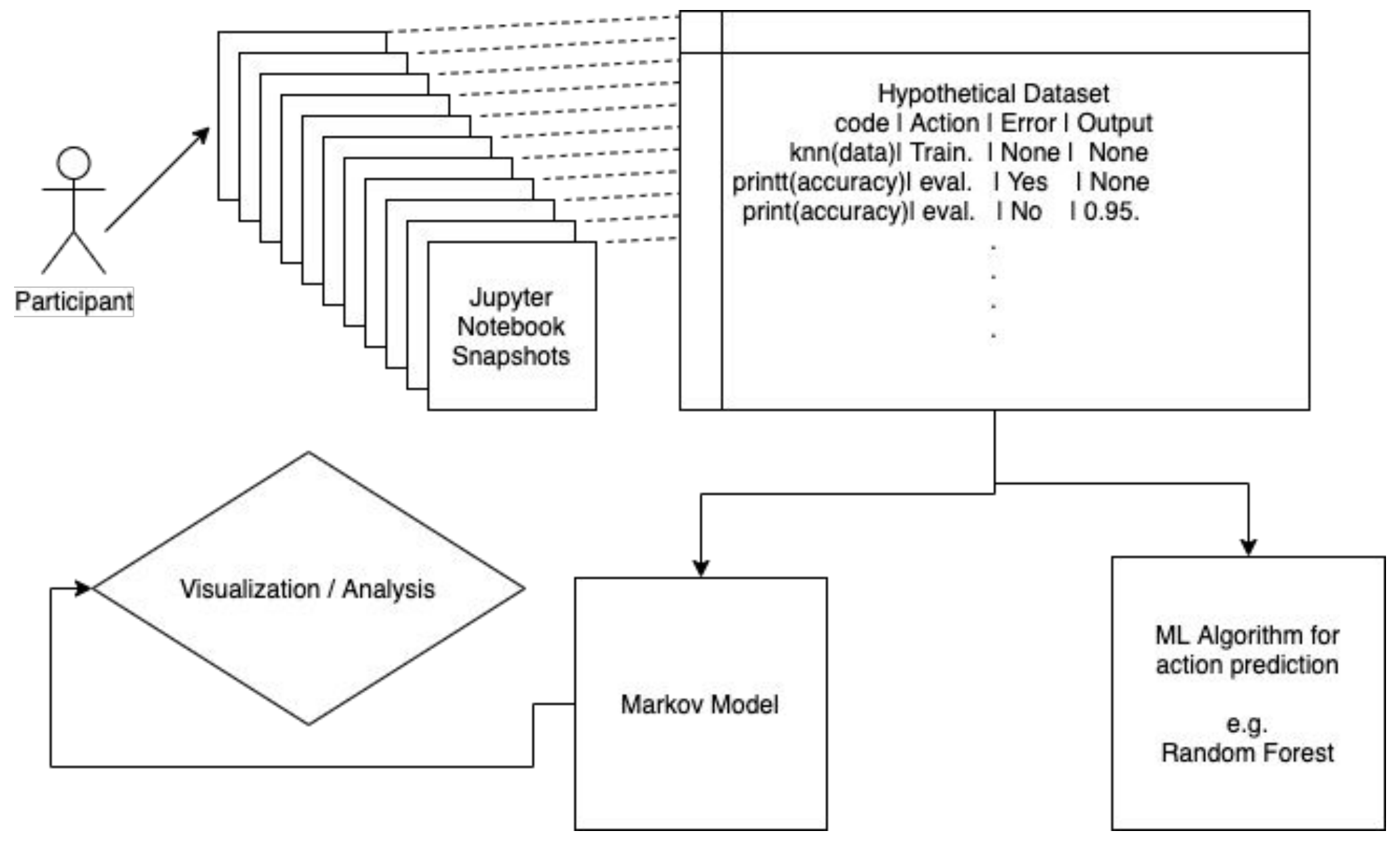
- Collected Jupyter notebook snapshots every few seconds
- Combined the notebooks and extracted the lines of code in order of temporal execution
- Also extracted error and output messages

Feature Engineering

- Labeled lines of code with the action taken (feature engineering, model training, evaluation, etc)
- Developed additional features for lines that led to errors or produced accuracy metrics, as well as features for previous several actions

Analysis

- Used a Markov model with first order assumptions to learn - probabilities of action transitions
- Examined the results to determine patterns in actions, used boosting methods with the engineered features to make action predictions



Hypothetical Dataset
code I Action I Error I Output
knn(data)I Train. I None I None
printt(accuracy)I eval. I Yes I None
print(accuracy)I eval. I No I 0.95.
.
.
.
.

Figure 1
A diagram of the pipeline used in our data extraction, engineering, and modeling process

VISUALIZATIONS AND ANALYSIS

Example Markov Model (Subject 2, Task 2)

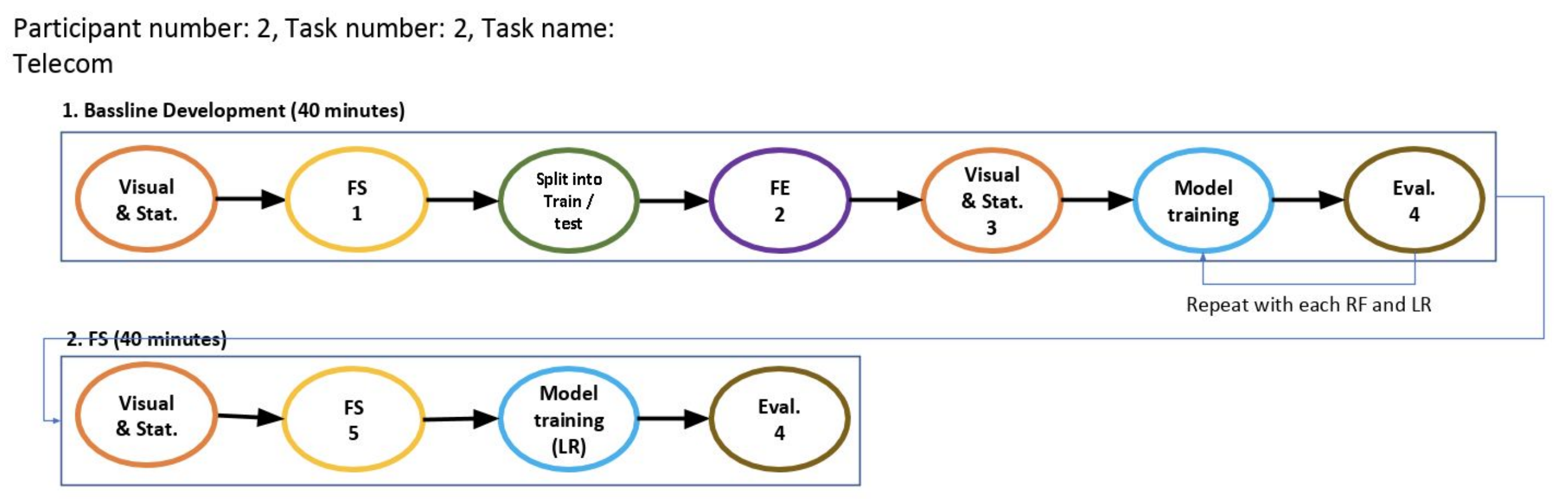


Figure 2
A diagram of the action processed used by this participant. We can see that the training-evaluation loop is supported by the Markov Model in **3**.

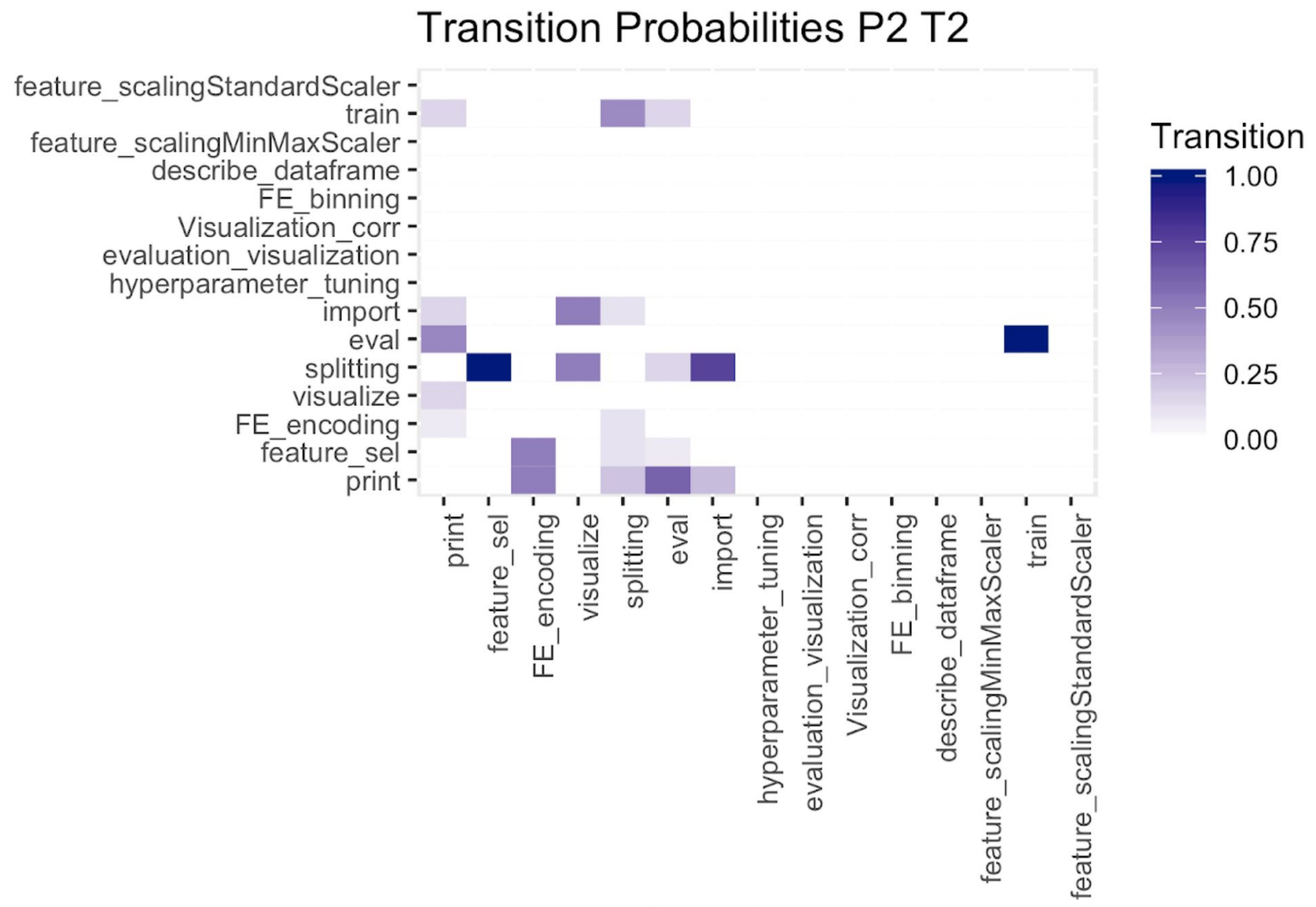


Figure 3
The X-axis on this heatmap represents the first action, and the Y axis represents the second. The color indicates the learned probability of the transition. This model was trained on all the participant data.

Prediction

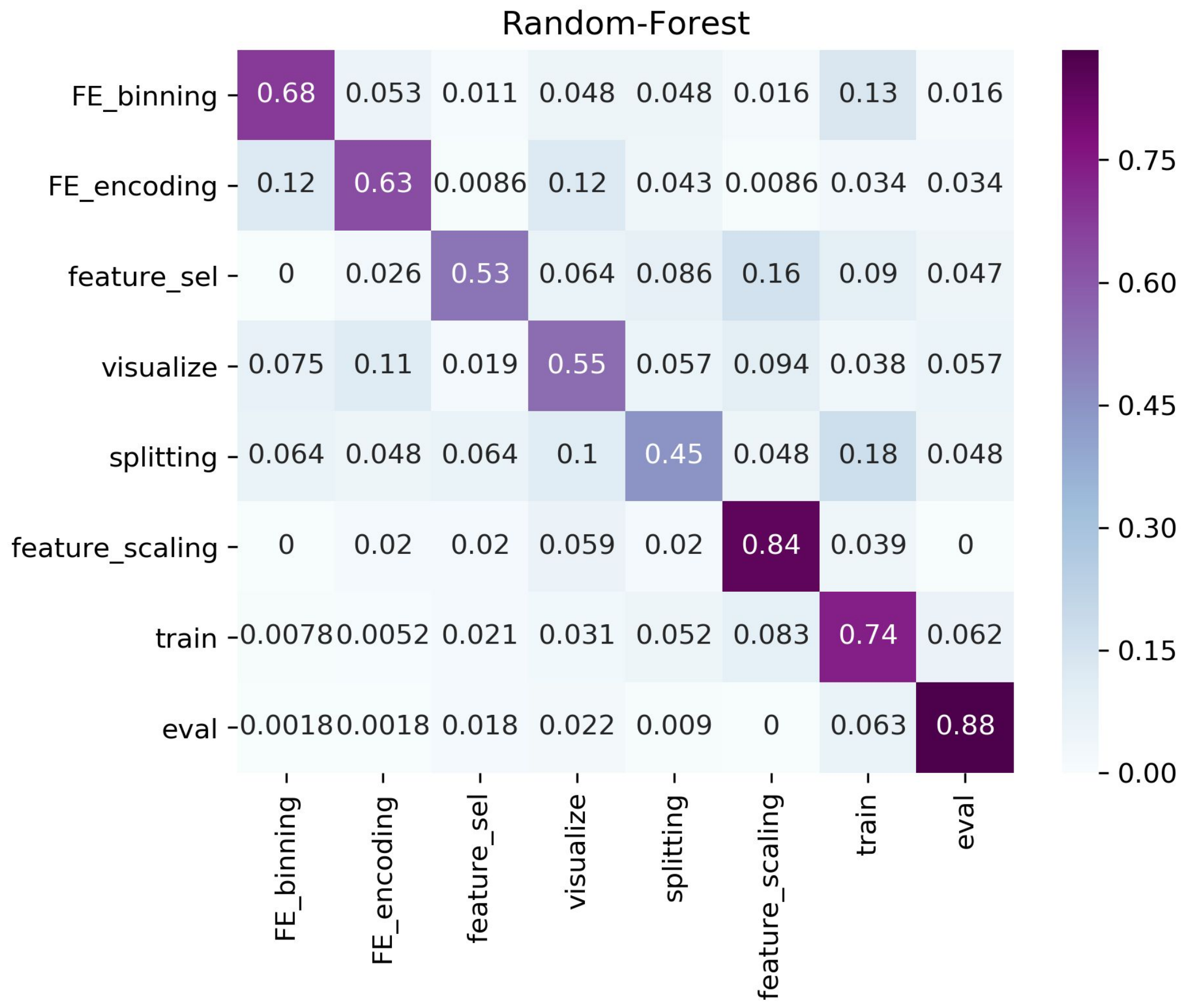


Figure 4
A plot of the leave one out accuracy and confusion of a random forest for action predictions, trained on engineered features of 3 previous actions, counter for occurrences of previous actions, and the 3 previous transitions. We see that the model performs well on training and evaluation, as those actions were common in sequence, but less well on actions like splitting, which led to several different next actions.

DISCUSSION

Findings

- Training and tuning was almost always followed with model evaluation
- Evaluation was often followed with training, indicating a repetitive pattern in problem solving
- Participants often visualized data and scaled features before training a model
- identified repeated execution of the same or similar lines as an indication of a bug in the participant's code.
- error detection will likely be a useful feature in the prediction of future actions.

Next Steps

We hope to employ the tagged actions, error detections, and outputs of participant's code as features to develop models to predict future actions of participants. This pipeline has been developed, and we are looking into new features that could be engineered to improve on the already promising results.

CONCLUSIONS

Through this research, we were able to develop and justify several insights into the process of solving a data science problem. Next steps include applying these findings in a predictive context, and engineering new features to improve the capabilities and reduce the biases in our own model.

Our findings here give us confidence in the ability to produce an agent that can provide recommendations, as we are currently able to explain and model several important trends in actions taken by the subjects.

Acknowledgements
I would like to thank Moshik Mash and the entire research team for all their guidance, teaching, and support.