Practicum Final Report

DSA-5900: Professional Practice (Summer 2022)

Credit Hours: four

Methamphetamine (Meth) Use Classification

By

David Nnamdi

Faculty Sponsor: Dr. Brian Fiedler

This proposal is presented in partial fulfilment of the requirements of the DSA 5900 course

Data Science and Analytics Institute

Gallogly College of Engineering

University of Oklahoma

Norman, USA

Contents

1.0 Introduction

In recent years, overdose deaths in the U.S. from stimulants or drugs other than cocaine has risen and Methamphetamine (Meth) abuse has been identified as the primary culprit for this rise (NIH, 2021). This has prompted an investigation into identifying meth usage amongst persons with substance abuse history.

Since 2014, the Substance Abuse and Mental Health Service Administration (SAMHSA), a section of the U.S. Department of Health and Human Services has published its annual National Survey on Drug use and Health (NSDUH). This survey typically measures illegal drug use, substance use disorder and treatment, and related users' mental health (Substance Abuse & Mental Health Data Archive, SAMHDA, 2022)

This project has been defined as an Imbalanced binary classification project with the end goal of predicting meth usage amongst people with recorded substance abuse history and understanding key factors contributing to these predictions in order to develop measures to mitigate meth abuse and likely overdose related deaths. This study will be focused on users surveyed in 2020.

2.0 Objectives

The objectives of this project as agreed by key stakeholders are:

- Feature selection to identify key predictors out of over 2000+ features
- Explore different sampling techniques for dealing with data imbalance and its effects on model performance
- Create a well-tuned Machine Learning model(s) that classifies an adult with substance abuse history as a Meth/Non-meth user
- Define Simple rules for meth usage identification amongst persons with substance abuse history

Personal learning objectives include understanding interdependency of substance abuse, understanding impact of economic factors on drug abuse and utilizing AutoML tools for parallel model training and evaluation.

3.0 Data

3.1 Ingestion

The data was obtained from 2020 NSDUH detailed tables published on the SAMHDA website here. Data was stored as tab delimited file which and was initially processed using MS excel into a csv file for further use.

Initial Dataset consisted of 32893 row entries and 2889 columns with a total size of over 725 megabytes. Each row entry corresponded to an individual user survey and each column represents answers to a specific question in the survey. These answers are either entered directly by users, logically assigned based on prior input, or imputed based on NSDUH predefined modified predictive mean Neighborhood, *modPMN* (SAMHDA, 2022).

3.2 Exploration

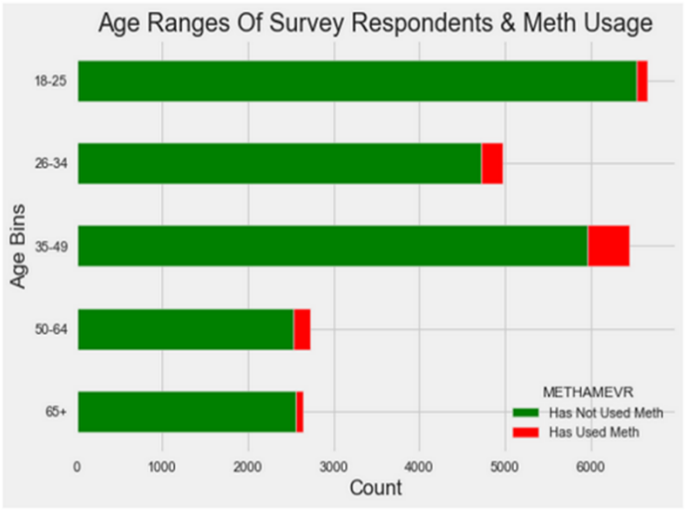
The initial dataset consisted of mostly of integer and float variables as with a single object type column containing the date the table was compiled. NSDUH had intentionally represented answers to questions (user specific, logically assigned, and imputed) with numerical variables in an attempt to standardize interpretation of answers. The table below details a high-level summary of numerical variables and their implied meanings:

Table 1: NSDUH Standard code conventions (SAMHDA, 2022)

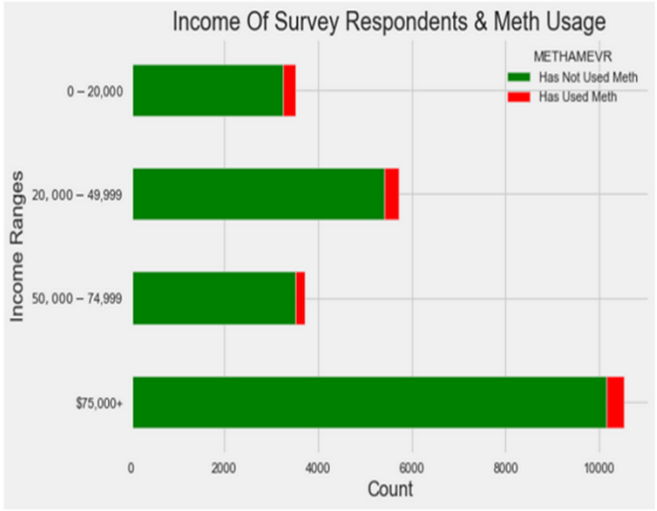
USER INPUT GENERATED	
91 or 991 or 9991	NEVER USED [DRUG(s) OF INTEREST]
93 or 993 or 9993	USED [DRUG] BUT NOT IN THE PERIOD OF INTEREST
94 or 994 or 9994	DON'T KNOW
97 or 997 or 9997	REFUSED
98 or 998 or 9998	BLANK (i.e., not answered; not asked the question)
99 or 999 or 9999	LEGITIMATE SKIP
LOGICALLY ASSIGNED	
81 or 981, 9981	NEVER USED [DRUG(s) OF INTEREST] Logically assigned
83 or 983, 9983	USED [DRUG] BUT NOT IN THE PERIOD OF INTEREST Logically assigned
85 or 985, 9985	BAD DATA Logically assigned (i.e., usually inconsistent with other data)
89 or 989, or 9989	LEGITIMATE SKIP Logically assigned

The data in table 1 above is not a comprehensive summary of all the numerical encodings of user answers for all sections. Sections that deal with user demographics, mental health and other specific questions have separate encodings; user age is also either entered as actual age or is binned and assigned a specific encoding.

In figure 1 we explore some demographics to understand age ranges and income of survey respondents and how meth usage is distributed across them. From 1a, we can easily observe that by count, most meth users are middle aged, between ages 35-49, with a sizable portion in their late twenties to mid-thirties. Income does not seem to play a very big role on meth usage as by raw count users are roughly evenly distributed across ranges (figure 1b).



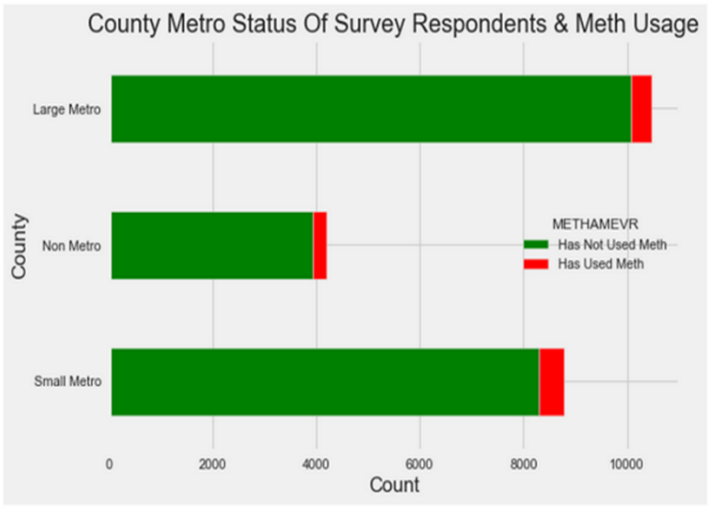
(a)



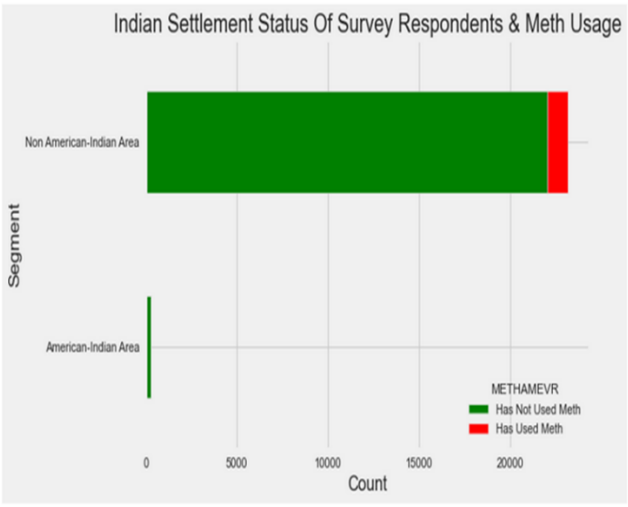
(b)

Figure 1: (a) Meth Usage by Age Group (b) Meth Usage by Income

In figure 2a and 2b we assess meth usage by county metro status and tribal land settlement. By raw count, it can be observed that a slightly larger number of meth users are in small metros, this number is smaller than meth users in large metros and non-metros combined. Meth usage is also only recorded in non-tribal lands – this may be biased to study sample and may not fully represent the larger population.



(a)



(b)

Figure 2: (a) Meth Usage by County Metro (b) Meth Usage in Tribal vs non-Tribal settlements

3.3 Preparation

3.3.1 Attribute Selection

A detailed review of the NSDUH codebook revealed that a lot of the questions asked in the survey were redundant, and the same information contained in several questions could be adequately represented by answers in one question. A basic example of these redundant questions related to recency of cocaine abuse includes:

- AGE WHEN FIRST USED COCAINE
- YEAR OF FIRST COCAINE USE
- # DAYS USED COCAINE PAST 30 DAYS
- BEST ESTIMATE #DAYS USED COCAINE PAST 30 DAYS

Color coding indicates similarity of both questions.

With knowledge of the above, some variables were eliminated entirely. In addition, the codebook was thoroughly reviewed and each question under different sections ranging from substance abuse to demographics to mental health were reviewed and specific questions (columns) were selected for further data preparation. This step reduced the total number of columns selected for further analysis from 2889 to 158 (including resultant feature to be predicted)

Note that all columns related to meth usage was removed to ensure no prediction bias.

3.3.2 Row level sub-setting

The following rows were dropped from the dataset:

- All rows where the answer to the question "EVER USED METHAMPHETAMINE" was not explicitly "Yes" or "No"
- All rows of all columns where code is equal to 94, 98 or 85 (refer to table 1 for details)

3.3.3 Factor Collapsing

Several columns contained more than ten distinct answer types – some of the standard convention code type as explained in table 1, others of unique encodings. Having reviewed the codebook, some of the data was collapsed into similar groupings as described below:

- User answers with code 91, 93, 81 and 83 are assumed to be same the encoding for the answer "No"
- User answers with code 99 and 89 are collapsed into a single factor 'skip'

Columns with AGE (user age or age of first substance abuse) had several distinct values and had to be collapsed as follows:

- For user Age, the following age groups were adopted:
    - 18-25yrs
    - 26-34yrs
    - 35-49yrs
    - 50-64yrs
    - 65+yrs
- For all age data related to recency of first use, the collapsed age groups were adopted:
    - Less than 18
    - 18 or older

3.3.4 Resulting Dataset

The above data preparation steps resulted in a dataset with 23454 rows and 158 columns with a total of 22306 non-meth users and1148 meth users, representing a class imbalance of 95:5.

The class imbalance of the number of meth users vs non-meth users is illustrated in the plot below:

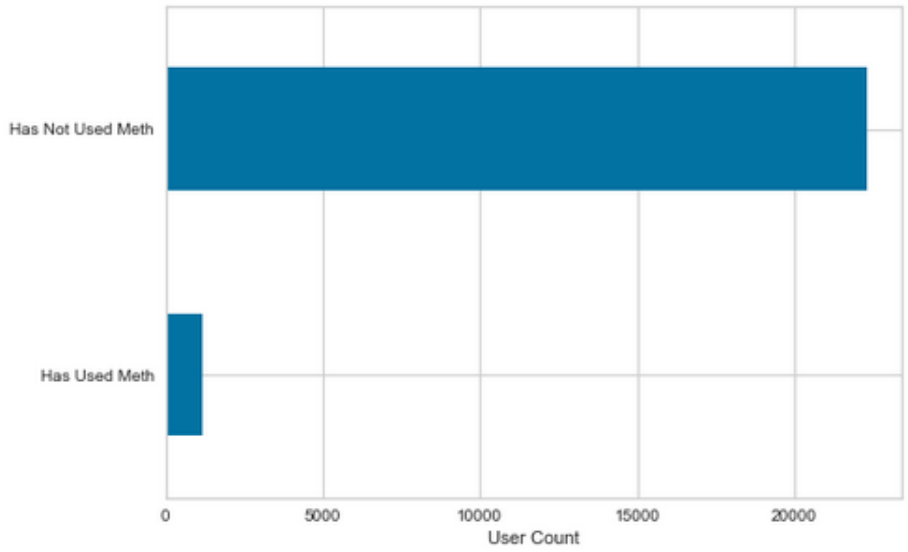


Figure 3: Methamphetamine usage amongst study respondents

The last step of the data preparation phase involved label encoding. Attributes may have been given numerical encodings by NSDUH, but they are categorical in nature and by description and were treated as such. Given that most of the data had now been collapsed into smaller factor levels, OneHot Encoding was used to generate the dataset to be used for modeling.

4.0 Methodology

4.1 Techniques

The first stage of the modeling involved defining representative samples for modeling. When dealing with imbalanced classification problems, the training data should be sampled such that there are similar number of positive and negative class input. This strategy helps improve model performance and generalization.

Three distinct types of sampling techniques were explored to deal with data imbalance. They are:

- Under sampling (the majority class, i.e., non-meth users)
- Oversampling (the minority class, i.e., meth users)
- Synthetic Minority Oversampling Technique (SMOTE)

The aim of trying these different sampling techniques was to assess model performance based on sample technique chosen. The modeling was done with tree-based ensembles which is a form of supervised learning. Several decision tree-based ensembles exist and rely on boosting or bagging of decision trees to improve predictive power.

In Boosting ensembles, parallel decision trees are built with random samples of features and each tree makes a prediction on test results with the most frequently predicted result being assigned as the final value. In contrast, building Bagging ensembles involves sequential training of decision trees and each tree is trained to reduce the misclassification rate of the previous tree.

For this study, the following tree-ensemble methods were evaluated:

- Random Forests (rf)
- CatBoost classifier (CatBoost)
- Gradient Boosting Classifier (gbc)
- Extreme Gradient Boosting (xgboost)
- Extra Trees Classifier (et)
- Light Gradient Boosting Machine (LightGBM)

The focus on tree-based ensembles is due to high performance on feature selection. For example, one of the outputs of training a model using random forests are the variable importance scores (VIS). Features with high VIS scores indicate predictors with the highest influence on correct classification (Beattie & Nicholson, 2020).

4.1.1 Process Validation

Some of the steps outlined during the data preprocessing phase are based on discussions with Dr. Matt Beattie who previously conducted a similar study feature selection for Heroin use prediction (Beattie & Nicholson, 2020).

Discussions with my Faculty Sponsor, Dr. Brian Fiedler has guided the project objectives and choice of supervised models. Dr. Brian's steer is focused on simplicity in modeling while emphasizing on extracting key rules that can aid in identification of likely meth users.

4.2 Procedure

The modeling was implemented using an open-source library called PyCaret, which is a Python equivalent for the Caret Package in R. The core reason for using this library is the ability to train several models in tandem and generate tabular comparisons of performance. Built in PyCaret functions also allowed for advanced plotting of several metrics to evaluate model performance and tune models with a defined target metric. Note that for most models, the default background framework being utilized for model building was the Scikit-learn library which is one of the most popular machine learning libraries available in the industry today.

In building the models for each training sample, stratified k-fold cross-validation (k=10) was used. This was done to ensure class distribution was maintained during splits and overfitting issues were reduced. The pre-processed data from the data preparation phase was split into 70-30 train test split, with a predefined random seed to ensure results are reproducible.

To evaluate performance of this binary classification model, three key metrics were of interest – Area Under Precision-Recall Curve (AUPRC), F1-score and the Recall (or sensitivity).

AUPRC is defined as the name implies, the area under a plot of precision vs recall. and may be utilized for evaluating models with heavily imbalanced datasets and is more informative than the ROC plot (Saito & Rehmsmeier, 2015)

where

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

F1-score is a measured balance of precision and recall and takes the harmonic average of both scores. F1-score is a variant of the more general F-beta score where Beta = 1.

$$F_\beta = \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (3)$$

In medical fields, reducing false negative is more important than reducing false positives. This convention is mostly adopted to reduce misdiagnosis while still having confidence in test coverage for true positive cases. Since early and accurate detection of Methamphetamine abuse is key to reducing overdose related deaths, some emphasis was placed on achieving models with high recall.

I was responsible for the entire data exploration, preparation, modeling, performance evaluation of models and key skills such as data understanding, visualization, sampling, stratified splitting, and cross validation techniques from the Advanced data analytics course formed the foundation of this analysis.

5.0 Results and Analysis

5.1 Modeling Results

5.1 Under-Sampling Results

Using the Scikit-learn under sampling class wrapper in PyCaret, a 50:50 balanced training dataset was extracted for model training. This was achieved by sampling the majority class (non-meth users) such that the sample matched the total number of minority class (meth users) in the training dataset. Generally, under sampling may lead to information loss and reduced model performance accuracy since you are training your model on a smaller sample size. Model generalization issues may also occur.

Table 2 below illustrates the model performance based on several metrics pre-built into PyCaret. The metric used to rank the table below is the AUPRC (alias APC). Note that metrics are averages evaluated on validation set from stratified k-fold cross validation.

Table 2: Under-sampling model performance metrics pre-tuning

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	APC	TT (Sec)
catboost	CatBoost Classifier	0.8770	0.9468	0.8817	0.2635	0.4054	0.3586	0.4412	0.5468	1.3660
rf	Random Forest Classifier	0.8679	0.9441	0.8921	0.2508	0.3913	0.3426	0.4307	0.5160	0.0890
gbc	Gradient Boosting Classifier	0.8734	0.9410	0.8753	0.2570	0.3971	0.3493	0.4324	0.5119	0.1850
lightgbm	Light Gradient Boosting Machine	0.8698	0.9387	0.8701	0.2505	0.3887	0.3401	0.4240	0.5117	0.1360
et	Extra Trees Classifier	0.8678	0.9441	0.8844	0.2492	0.3887	0.3398	0.4269	0.5002	0.0830
xgboost	Extreme Gradient Boosting	0.8642	0.9368	0.8804	0.2435	0.3813	0.3317	0.4195	0.4938	0.3640

Based on initial model training, it can be clearly observed that CatBoost model performed the best with an ***AUPRC of 0.5468***. However, since these models were base models without any hyper-parameter tuning, an attempt to improve model performance was made by tuning the top three models. Post tuning, the random forest model had the overall best performance on test data with an ***AUPRC of 0.6056*** an ***F1-score of 0.4657*** and ***recall of 0.9*** as illustrated in table 3 below

Table 3: Tuned Random Forest classifier performance metrics on test data

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	APC
0	Random Forest Classifier	0.8914	0.9545	0.9000	0.3142	0.4657	0.4206	0.4935	0.6056

Figure 4 illustrates some standard performance plots including the precision-recall curve, a confusion matrix indicating only *37 false negative* predictions out of 370 highlighting the model's high recall. It can be observed that there are a sizable number of *false positive predictions (727)* indicating low model precision, however, in comparison to the correctly predicted negative classes, it is a relatively small number (~11%).

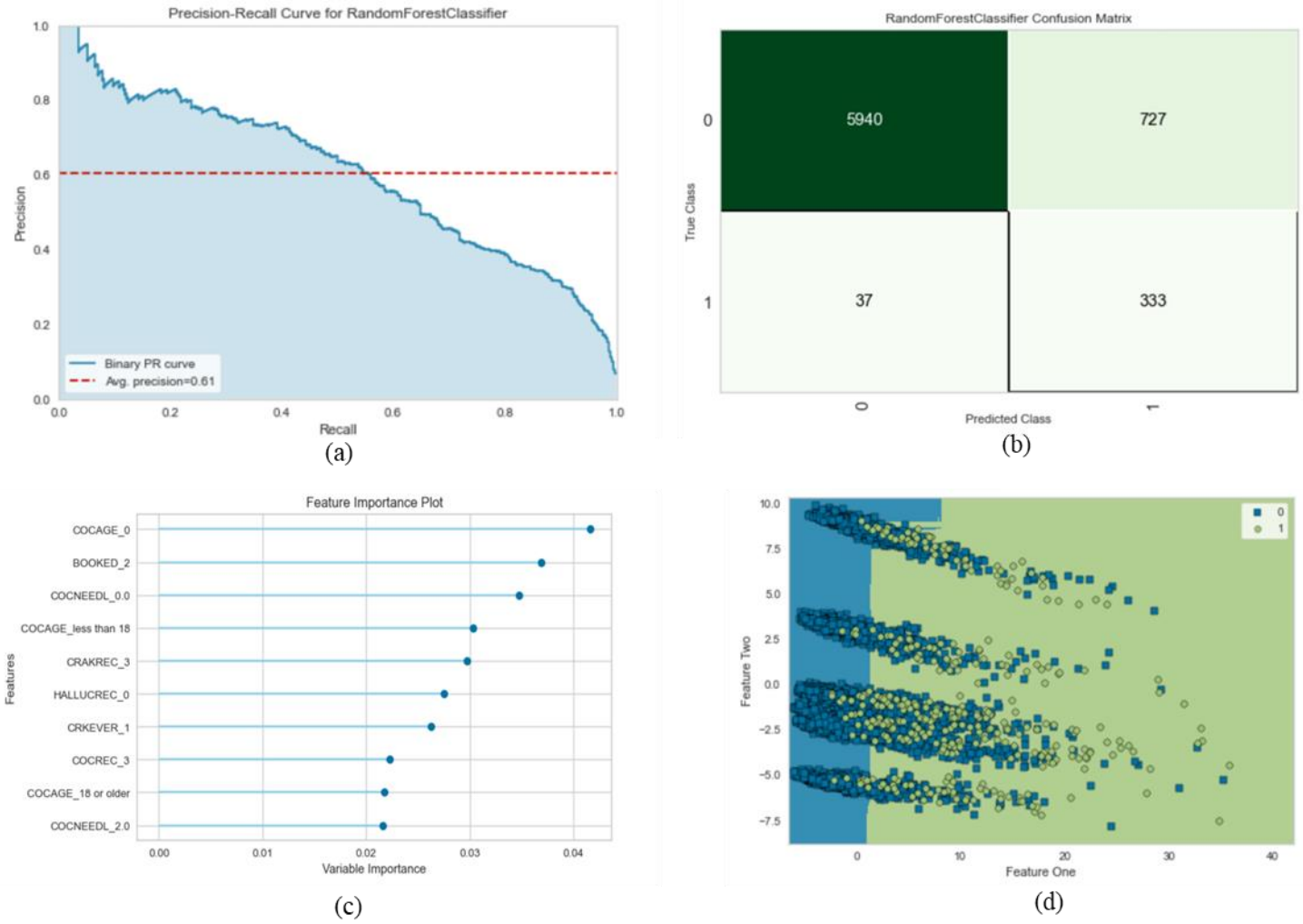


Figure 4: Random Forest model evaluation plots. (a) P-R curve (b) Confusion matrix (c) Feature Importance plot (d) Decision Boundary plot

Figure 4c highlights the ranked feature importance of the model with ***Age of first cocaine use*** and ***booking/arrest history*** being the top two features. Figure 4d illustrates the 2D decision boundary of the random forest model and emphasizes overlap of classes in the green region representing meth-usage. This overlap is responsible for the poor precision of model.

5.2 Over-Sampling Results

A similar modeling approach was taken for the training samples generated by oversampling. During oversampling, duplicates of the minority class (meth users) are made to balance the majority class samples. This method ensures minimal information loss but may cause model overfitting issues of the minority class.

Table 4 below shows the performance of key tree-ensemble models on the oversampled training dataset:

Table 4: Over-sampling model performance metrics pre-tuning

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	APC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8926	0.9470	0.8560	0.2886	0.4313	0.3879	0.4587	0.5612	3.1820
et	Extra Trees Classifier	0.9592	0.9369	0.1979	0.7780	0.3132	0.2998	0.3775	0.5489	1.3480
rf	Random Forest Classifier	0.9598	0.9417	0.3123	0.6696	0.4241	0.4061	0.4390	0.5486	0.9130
catboost	CatBoost Classifier	0.9470	0.9380	0.5924	0.4590	0.5156	0.4882	0.4935	0.5295	3.5080
lightgbm	Light Gradient Boosting Machine	0.9335	0.9396	0.6979	0.3897	0.4994	0.4670	0.4901	0.5271	0.3190

Here the Gradient Boosting Classifier performs best based ***AUPRC metric (AUPRC: 0.5612).***

An attempt to tune top three models still left the gbc as the best performing model on test data with an ***AUPRC of 0.6150*** an ***F1-score of 0.4776*** and ***recall of 0.9081***. as illustrated in table 5 below. This performance is better than the results of the models trained with under sampled data and reflects information loss and poor generalization highlighted in previous section.

Table 5: Tuned Gradient Boosting classifier performance metrics on test data

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	APC
0	Gradient Boosting Classifier	0.8956	0.9569	0.9081	0.3240	0.4776	0.4337	0.5056	0.6150

Reviewing the confusion matrix in figure 5b below we see that the *false negative predictions were 34* and *false positives 710*, a reduction from values seen in the previous section. In figure 5c we see that again ***Age of first cocaine (COCAGE)*** use is the most important feature followed loosely by ***recency of Hallucinogen use (HALLUCREC)*** and then ***booking/arrest history (BOOKED)*** coming in at third place. This indicates some sort of feature consistency across different modelling/sampling techniques.

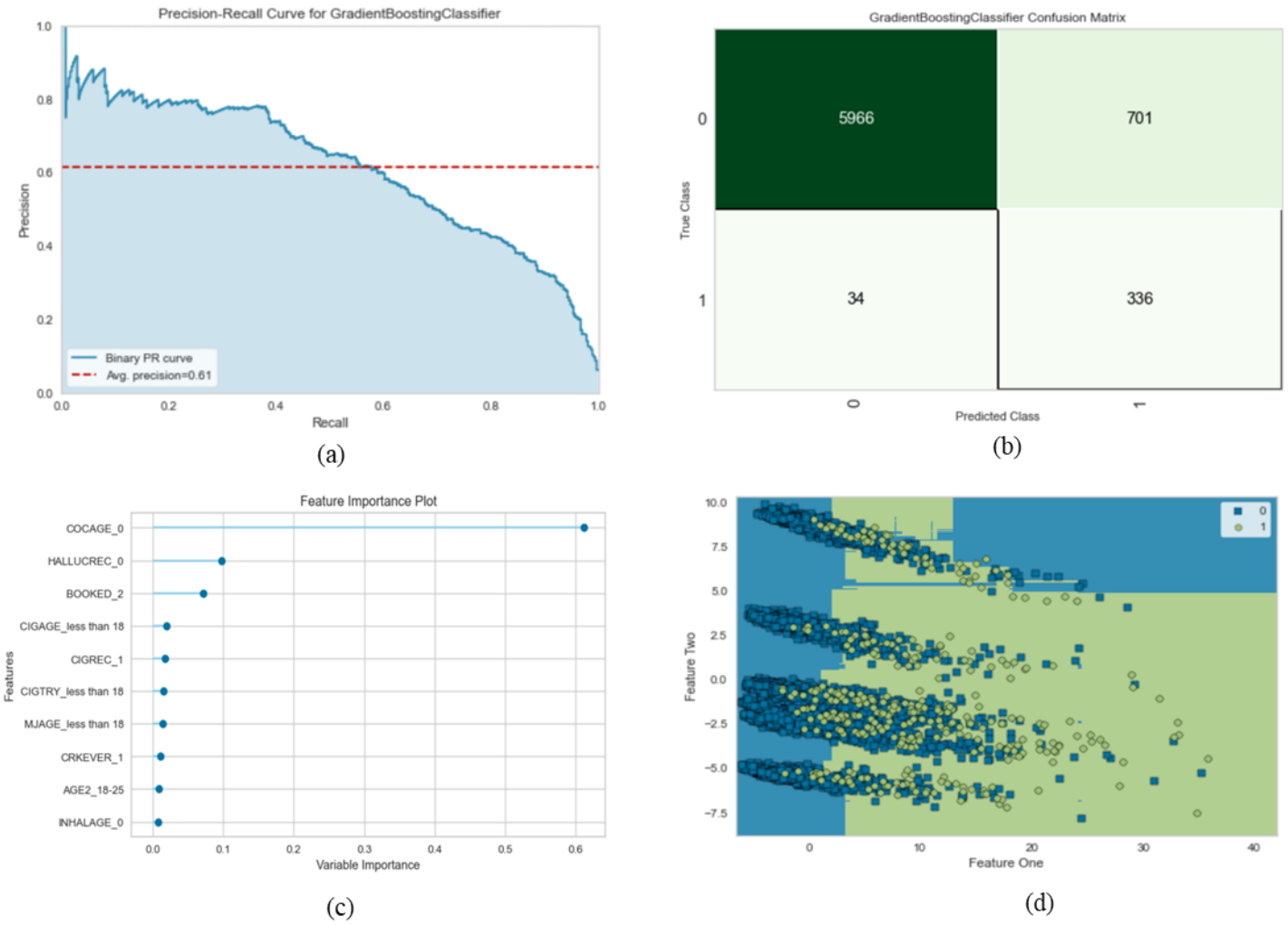


Figure 5: Gradient Boosting Classifier model evaluation plots. (a) P-R curve (b) Confusion matrix (c) Feature Importance plot (d) Decision Boundary plot

One can immediately observe that in figure 5d, additional decision boundaries were added for the gbc model, and this is what lead to better classification metrics.

5.3 SMOTE-Sampling Results

A final attempt to balance the model was done using the Synthetic Minority Oversampling Technique (SMOTE). (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) defined the technique where samples along a line connecting a random sample to one of its k nearest neighbors are synthetically generated and used to boost sample size distribution of minority class. This method has been adopted largely in the Artificial Intelligence and Machine Learning space because the synthetic samples generated are plausible based on closeness to other minority samples.

Table 6 below shows the performance of key tree-ensemble models on the SMOTE training dataset.

Table 6: SMOTE model performance metrics pre-tuning

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	APC	TT (Sec)
catboost	CatBoost Classifier	0.9605	0.9457	0.3893	0.6464	0.4829	0.4637	0.4814	0.5587	20.9740
gbc	Gradient Boosting Classifier	0.9594	0.9477	0.4574	0.6011	0.5167	0.4959	0.5024	0.5575	5.5560
lightgbm	Light Gradient Boosting Machine	0.9589	0.9461	0.3753	0.6175	0.4646	0.4446	0.4607	0.5478	1.1170
et	Extra Trees Classifier	0.9591	0.9450	0.4137	0.6037	0.4893	0.4688	0.4788	0.5410	1.4170
rf	Random Forest Classifier	0.9589	0.9449	0.3098	0.6443	0.4162	0.3977	0.4275	0.5400	0.9310

The best performing model based on the initial training is the CatBoost classifier with an average AUPRC of 0.5587 from cross-validation.

On fine-tuning the hyper-parameters of the top models, we end up having the Light Gradient Boosting Machine as the best performing model on the test set with the highest recorded ***AUPRC: 0.6375*** and ***F1-score: 0.5714***. This model however has low recall (as seen in table 7 below) when compared to the best performing models for under-sampling and over-sampling and favors improvements in precision. This may not be especially useful for the purpose of meth usage classification.

Table 7: Light Gradient Boosting Machine performance metrics on test data

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	APC
0	Light Gradient Boosting Machine	0.9604	0.9595	0.5027	0.6619	0.5714	0.5511	0.5568	0.6375

Figure 6 shows LightGBM model performance and feature importance. The confusion matrix in figure 6b shows we have 184 out of 370 false negative classifications however the false positive classification is reduced to only 95.

We can observe a very different set of important features to what has been observed so far in the under-sampling and oversampling best performing models. In figure 6c, we observe that LightGBM ranks ***booking/arrest history (BOOKED)*** as the most important feature, followed by ***recency of cigarette usage (CIGREC)*** and ***age of first trying cigarettes (CIGTRY)***.

The decision boundary (figure 6d) formed by LightGBM for classification is similar to that generated by the random forest model trained on under sampled data. It contains low coverage for negative points (blue) that may lie in regions of dominant positive predictions (green). Note that negative implies non-meth usage and positive vice-versa

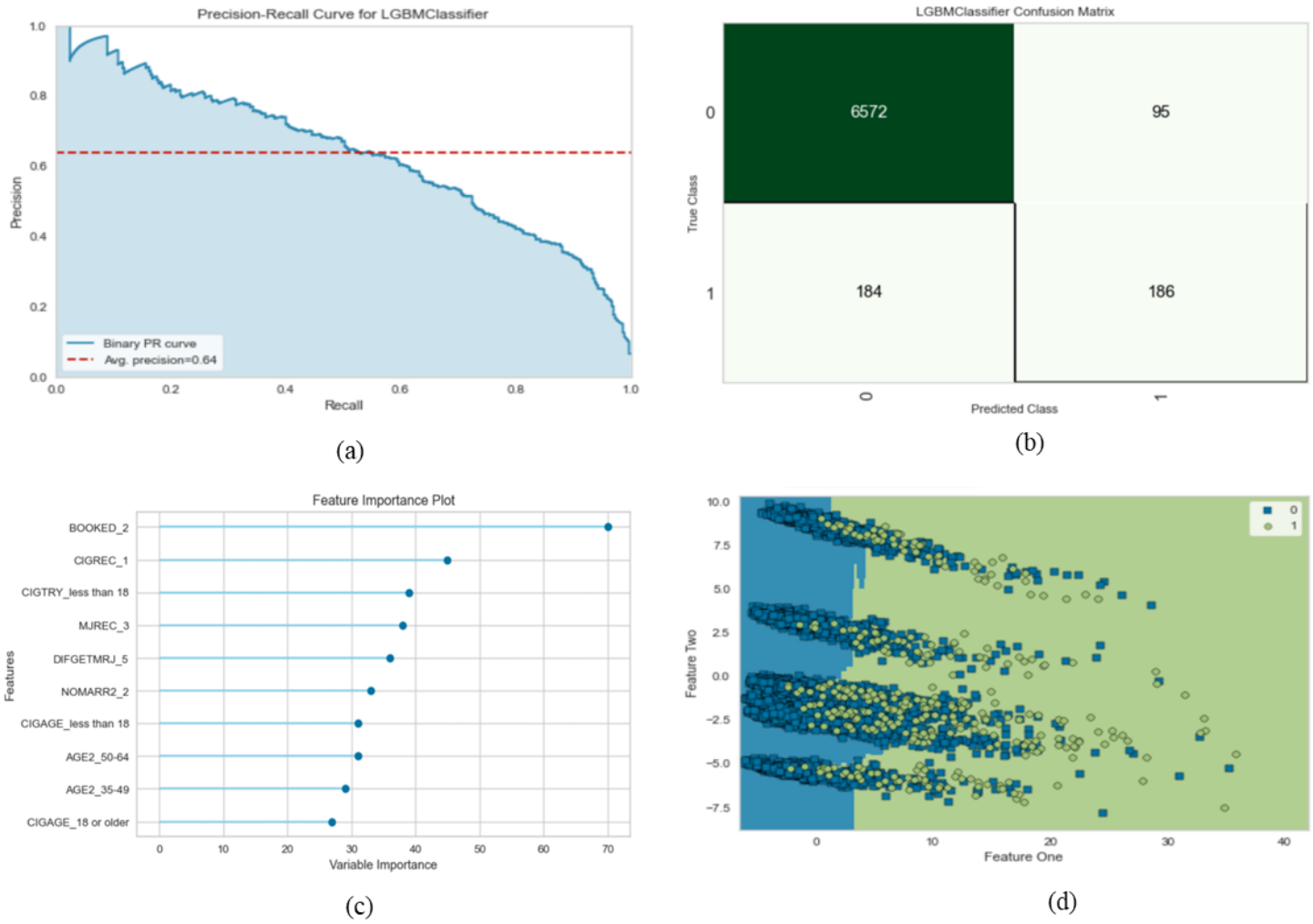


Figure 6: Light Gradient Boosting Machine model evaluation plots. (a) P-R curve (b) Confusion matrix (c) Feature Importance plot (d) Decision Boundary plot

5.2 Analysis

5.2.1 Evaluation & Final Model Selection

A comparison of the three best models trained using the three different sampling techniques are presented in figure 7. Based on AUPRC and F1-score alone, a bias selection of the LightGBM model may be made but considering the importance of reducing false negatives in our predictions, **the best model out of the three becomes the Gradient Boosting Classifier (gbc) trained on the oversampled data**. This model has the best recall value, and second best AUPRC and F1-scores of the three models.

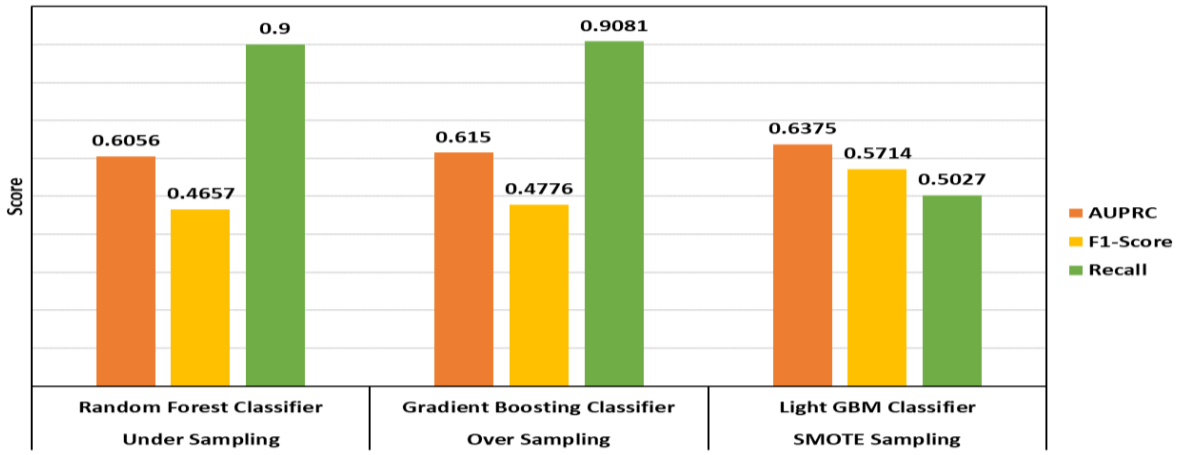


Figure 7: Best performing models performance comparison across different sampling techniques

Using some key features of the gbc model, an attempt was made to formulate some simple rules for guiding meth usage prediction. The features considered are:

1. Age of First Use of Cocaine (COCAGE)
2. Hallucinogen Use Recency (HALLUCREC)
3. Arrest History (BOOKED)
4. Age of Marijuana First Use (MJAGE)

5.2.2 Important Feature EDA

Figure 8 show meth usage for distinct groups based on age of first use of cocaine and marijuana. In figure 8a, it can be easily observed from the relative bar sizes that people who have never used cocaine are less likely to use meth, people that first used cocaine before 18 are more likely to use meth and people that used when 18 or older are somewhat in the middle.

In figure 8b, the highest likelihood of meth usage is associated with under-18 usage of marijuana. Almost the entire population of surveyed respondents that have never used marijuana never use meth.

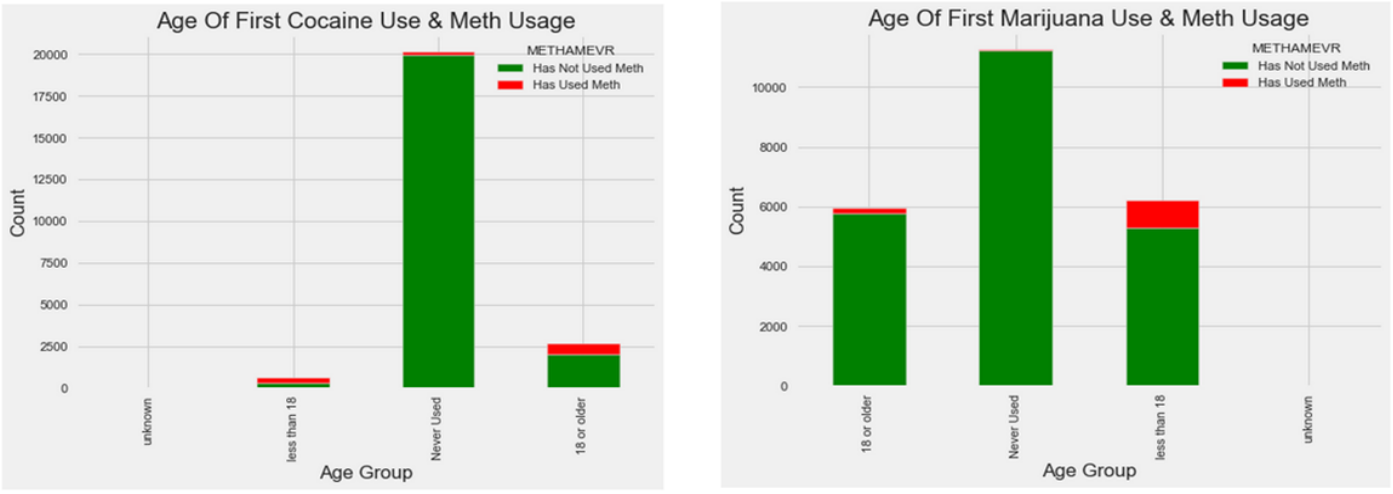


Figure 8: Meth Usage by (a) Age of first cocaine use (b) Age of first marijuana use

In figure 9, recency of hallucinogen use, and arrest history are compared to meth usage. We observe that respondents with hallucinogen use within and around a year have higher likelihood of meth usage and having an arrest history increases the likelihood of meth usage.

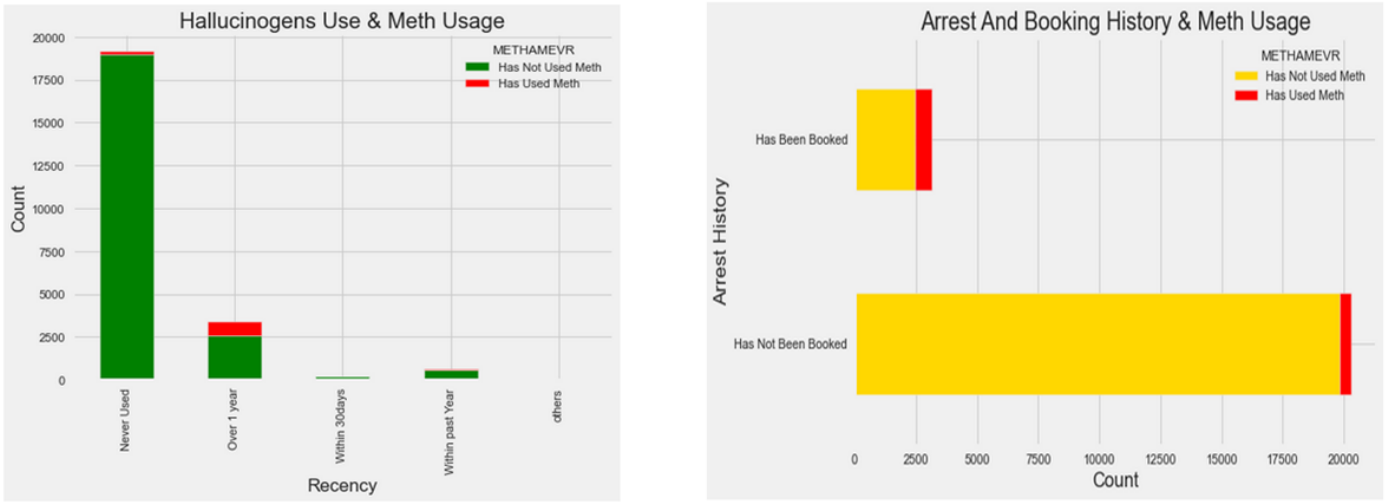


Figure 9: Meth usage by (a) Recency of Hallucinogen use (b) Arrest and booking history

5.2.3 Rule Formulation

The formation of simple rules using those four key features can be done in a variety of ways, one of which involves building a shallow depth decision tree with the features and another just based on logical interpretation of the EDA. An attempt was made to build a shallow decision tree was made, here depth = 3, and the resulting tree is shown in figure 10 below:

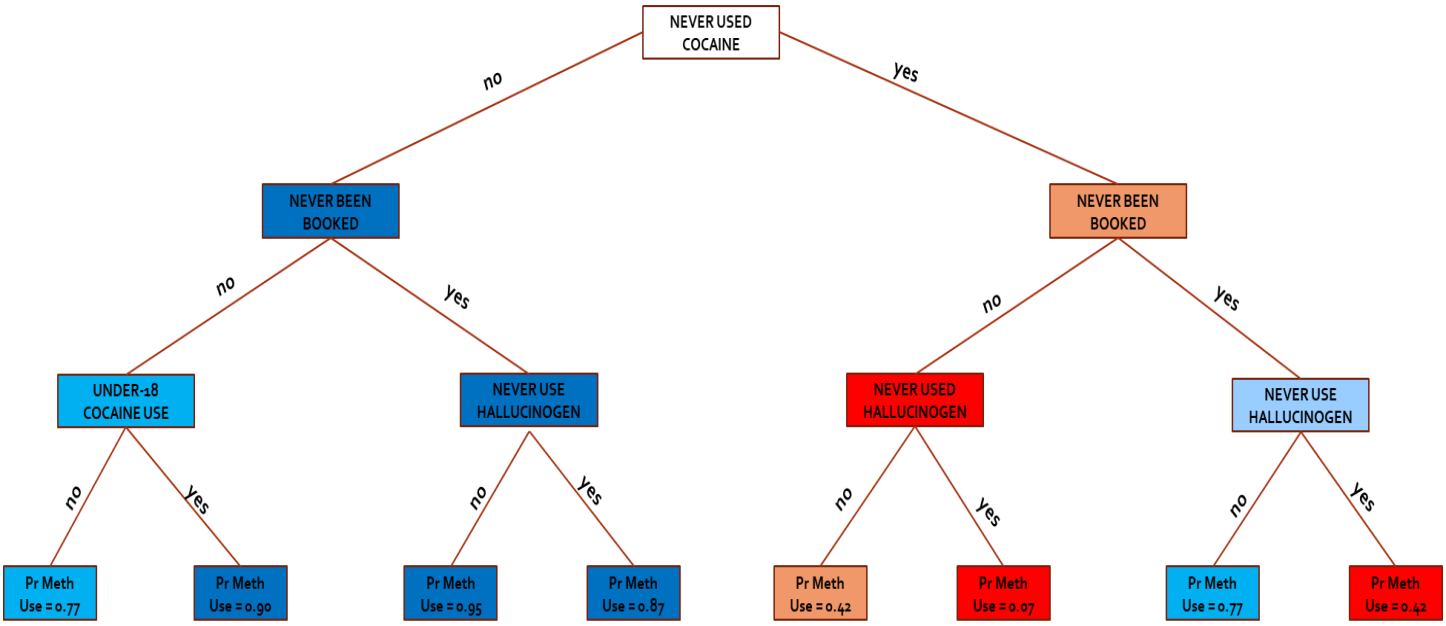


Figure 10: Shallow decision tree for rule formation (3 out of 4 key features are used)

Alternatively, for more intuitive rule formation the following rules are proposed:

1. A person with **early first Marijuana usage (<18 years)**, **Adult Cocaine usage** and **Hallucinogen use within the past year** and **has arrest history** is **more likely** to delve into meth usages in their mid-30's to late 40's
2. A person with **no history of cocaine or hallucinogen usage** and **adult usage of marijuana** with **no prior arrests** or bookings **less likely** to delve into meth usage

6.0 Deliverables

This study was designed to tackle an imbalanced classification problem related to predicting meth usage amongst persons with substance abuse history. By careful methodology design, the set project objectives were achieved.

So far, this study has demonstrated the effectiveness of oversampling and tree-based ensembles in imbalanced classification problems with our best performing model being the gradient boosting classifier ensemble. In building the model, a trade-off between precision and recall had to be made, with recall being a more principal metric for this classification as we aimed to minimize false negatives.

Through combination of model feature importance rankings and EDA, simple rules were formulated that can help the U.S department of health identify possible meth users and make efforts in reducing meth-related deaths

Underage Marijuana abuse has been linked to adult meth usage and while regulations already exist to forestall this from happening, the current widespread access of Marijuana and cannabis-based products in states where it is legal may create unwanted underage access. The U.S government may try to enforce stricter restrictions to ensure no underage marijuana abuse occurs.

For Further research on the topic, researchers may investigate societal influences that lead to meth usage

7.0 Self-Assessment

My most important learning on the project was downsizing feature set while still maintaining interpretability. This required me to ignore traditional dimension reduction techniques like PCA and focus more on the data understanding and interpretation. Other utilized data skillsets include data preparation (label encoding, sampling) and model evaluation (cross validation, visualization of decision boundaries, P-R curves).

Interdependency of cocaine, marijuana and hallucinogen use with meth usage was established and this was a primary personal learning objective. For this study I also had to learn how to write Python class wrapper functions for Scikit-learn to able to use certain metrics of interest in evaluating my model performance. I solidified my knowledge on AutoML use with focus on the PyCaret package.

References

Beattie, M., & Nicholson, C. (2020). Feature extraction for heroin-use classification using Imbalanced Random Forest Methods. *Substance Use & Misuse*, 123-130.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321-357.

NIH. (2021, 10 05). *Trends in U.S. methamphetamine use and associated deaths.* Retrieved from National Institutes of Health: https://www.nih.gov/news-events/nih-research-matters/trends-us-methamphetamine-use-associated-deaths

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One vol 10(3):e0118432*.

SAMHDA. (2022, 05 30). *NSDUH Codebook.* Retrieved from https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2020/NSDUH-2020-datasets/NSDUH-2020-DS0001/NSDUH-2020-DS0001-info/NSDUH-2020-DS0001-info-codebook.pdf

Substance Abuse & Mental Health Data Archive (SAMHDA). (2022, 06 01). *National Survey on Drug Use and Health (NSDUH).* Retrieved from https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001