Genomically Predicting Antibiotic Resistance in Gonorrhea

Introduction

With an estimated 1,568,000 new infections per year, gonorrhea causes a burden on the U.S. healthcare system that amounts to $271 M each year and resulting in health issues including infertility, ectopic pregnancy, and potentially increased risk of HIV contraction. This bacterial infection is commonly treated with antibiotics, but with the rise of antibiotic resistance, half of all infections now resist at least one treatment. In order to prevent further rise of antibiotic resistance and the creation of superbugs, DNA analysis presents an opportunity for the identification of effective treatments.

Using unitigs from genetic data in 12 studies that sequenced antibiotic-resistant strains, I created three models to predict the resistance of gonorrhea strains to azithromycin (recall = 0.68) , ciprofloxacin (recall =0.96), and cefixime (recall = 0.89) from the unitigs present in a strain. This process can be repeated for other antibiotics and other bacterial infections to provide targeted effective therapies to patients.

Data Wrangling

The raw data consisted of four files: one unitig file per antibiotic that contains DNA sequences and whether or not the strain contains said unitigs and another file containing all strains and their minimum inhibitory concentration (MIC) that measure the degree of resistance to azithromycin (azm), ciprofoxacin (cip), and cefixime (cfx).

I created the labels for each strain of resistant or sensitive using the MIC data. To do so, I converted and corrected non-numeric MIC data at the upper and lower bounds of tested concentrations. As MIC is determined by dilution, our safest assumption that twice the maximum concentration is the most accurate MIC, so 512 for >256 and 1024 for >=512. It is important to retain these values as they are the most resistant to the antibiotic. Similarly, for <=0.008, we will assume an MIC of one dilution further: 0.004, which is still far below most values and well within the sensitive range for the antibiotic. Once all data was in numeric form, I determined resistance using the following boundaries:

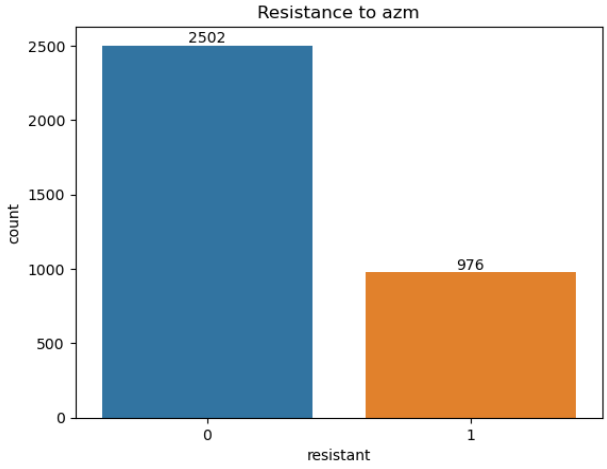
Antibiotic	Resistant (R)
azm	> 0.50 mg/L
cip	> 1.00 µg/mL
cfx	> 0.125 mg/L

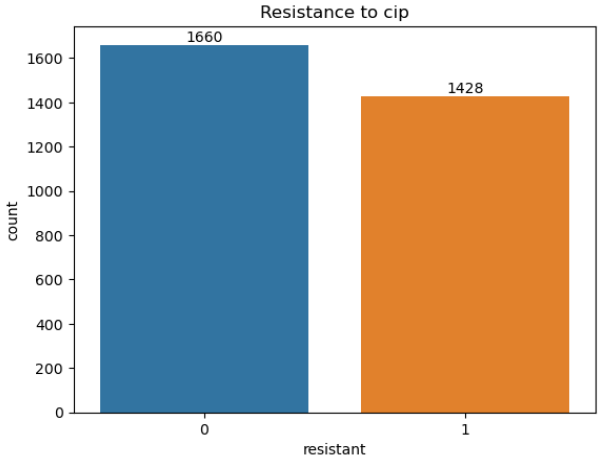
Finally I joined the tables to have three data sets that included the strains with their present unitigs and a labels column of resistant or not and dropped strains with missing labels as well as unitigs not present in any strains. The final datasets are shaped as follows:

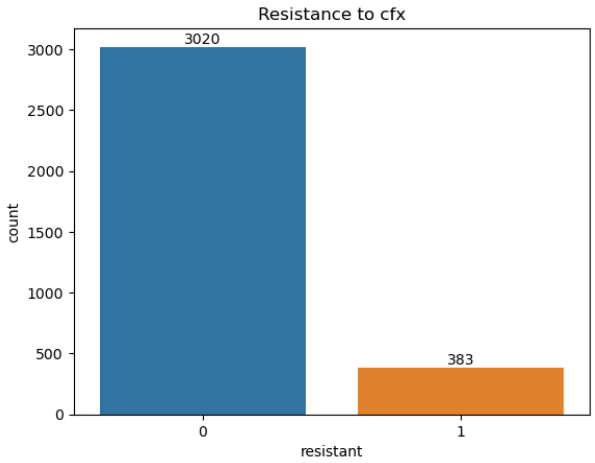
Antibiotic	Rows (number of strains)	Columns (number of unitigs)
azm	3478	492
cip	3088	8488
cfx	3403	363

Data Analysis

First I explored the prevalence of resistance to each antibiotic:

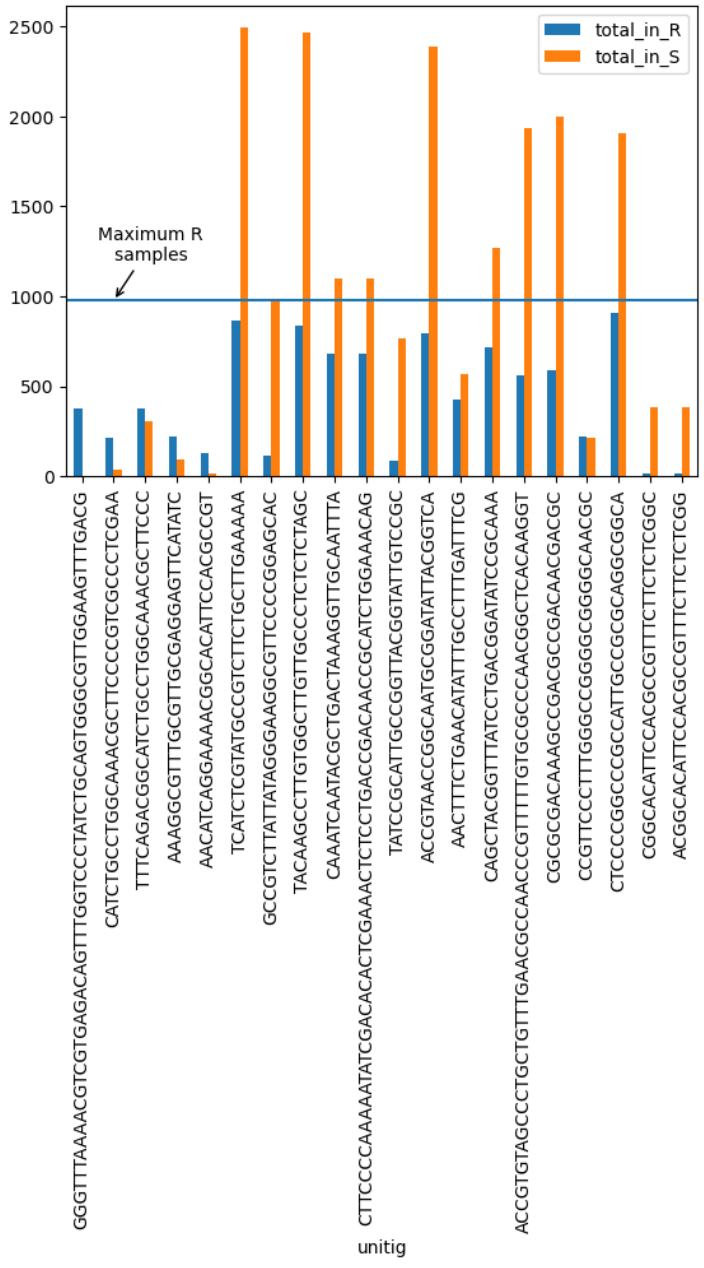




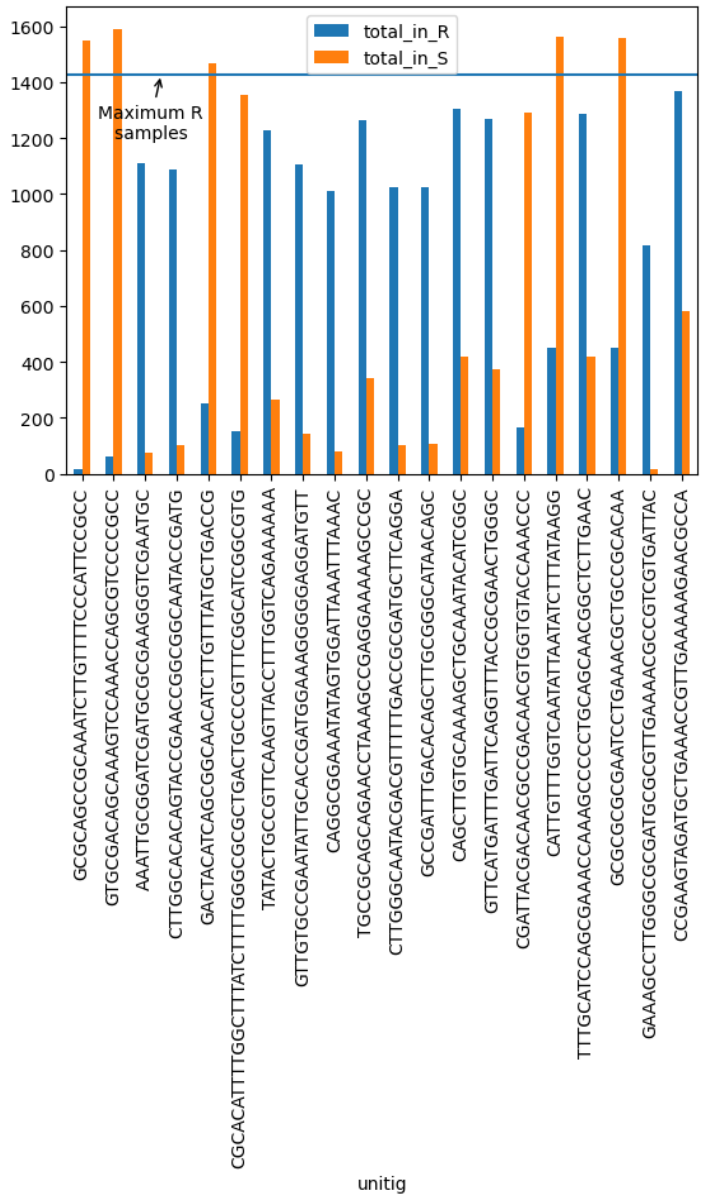


Cfx showed the largest class imbalance as only 11% of the strains were resistant, as it is the most recently used antibiotic. Cip resistance has the highest prevalence as it was a common primary treatment for years.

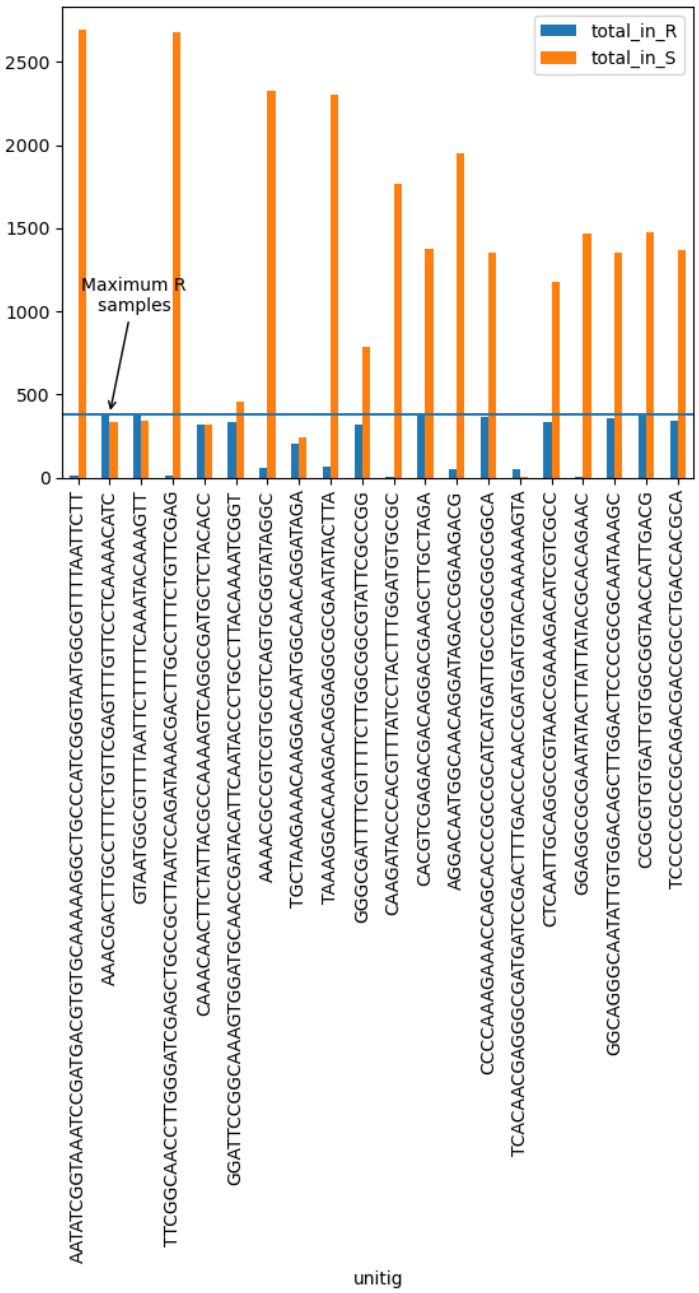
The unitig features in the data were binary with a 1 indicating the unitig is present in the strain and a 0 indicating its absence. With 300-8000 feature columns, I conducted Chi-squared analysis to determine features of importance in models.



Azm had the least correlated unitigs overall, but one unitig with a Chi-squared value of 1089 was present in 380 resistant samples and only 1 sensitive samples, indicating a potentially selective feature. Other unitigs showed more prevalence in exclusively sensitive strains as well. However, even in the more correlated features, a substantial amount of overlap appears.

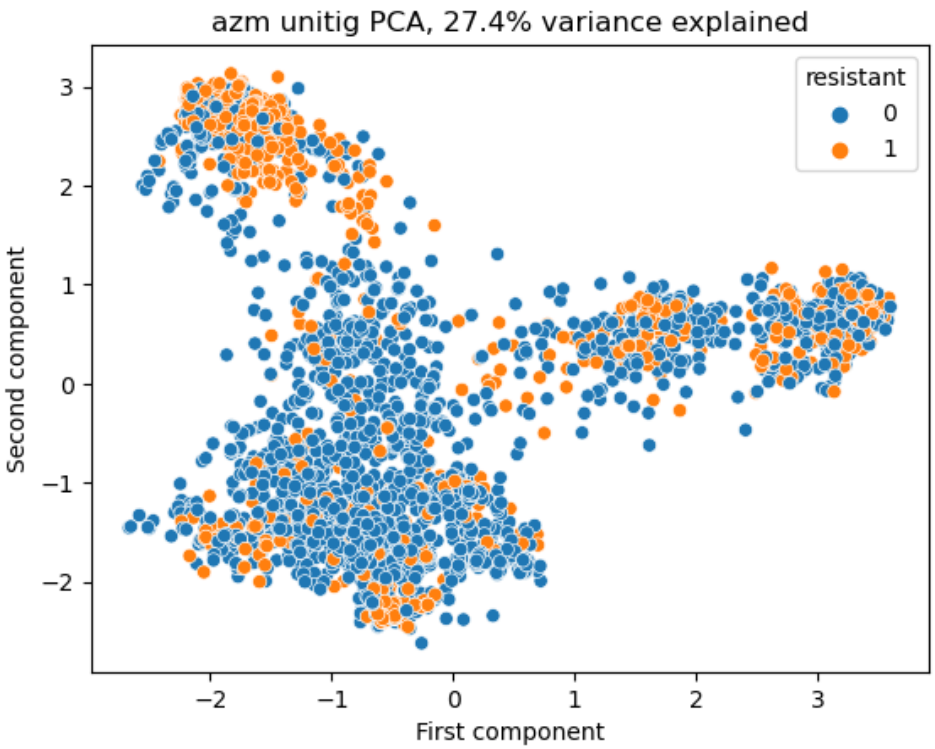


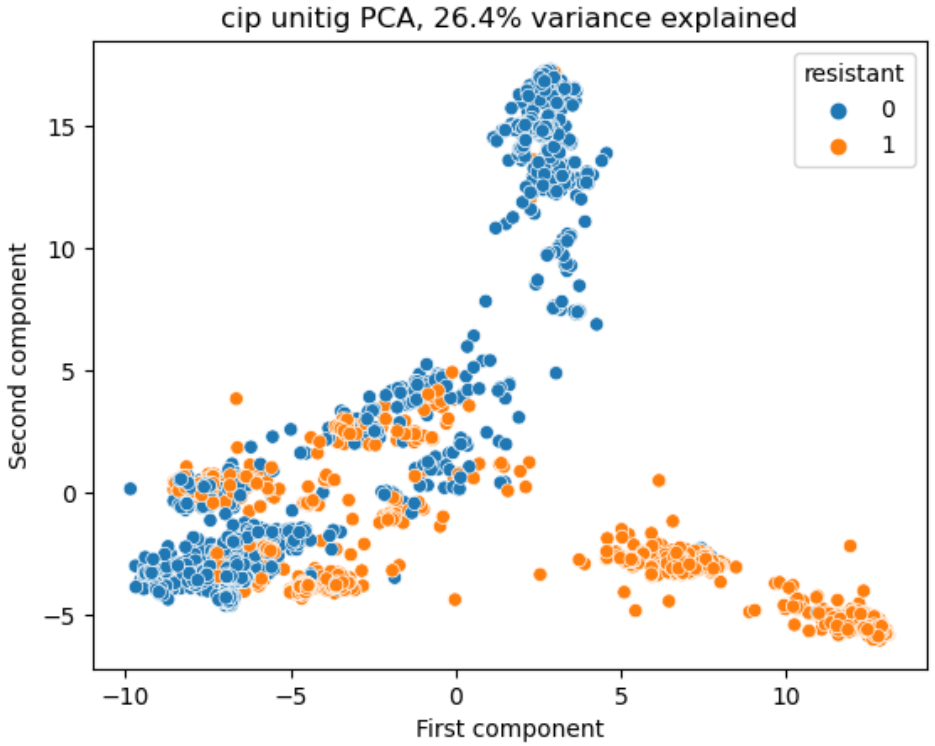
Cip, on the other hand, had Chi-squared values of up to 2618. There was clear separation within these unitigs of their presence in resistant or sensitive strains, providing promise for a selective model.

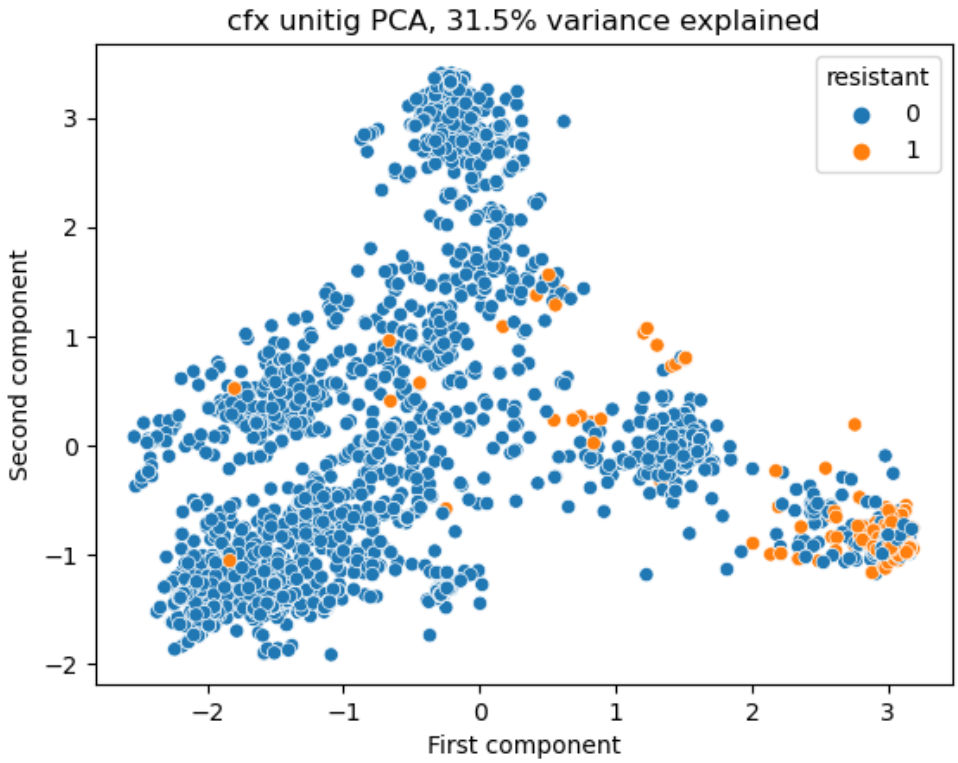


Despite having the most class imbalance, cfx had unitigs more correlated with resistance than azm, with 5 features having higher Chi-squared values than azm's most correlated unitig. Two of said unitigs are only present in 11 of the 363 resistant strains and another two are present in 375 of the resistant strains while being in 11% of the sensitive strains.

To test if clustering may reveal patterns with the data, I conducted PCA on the three datasets. However, the first two principal components only accounted for 26-31% of the variance in each case. Cip revealed the highest separation using PCA but still substantial overlap was noted between resistant and sensitive strains.


azm unitig PCA, 27.4% variance explained
resistant
0
1
Second component
First component
3
2
1
0
−1
−2
−2
−1
0
1
2
3


cip unitig PCA, 26.4% variance explained
resistant
0
1
Second component
First component
15
10
5
0
−5
−10
−5
0
5
10

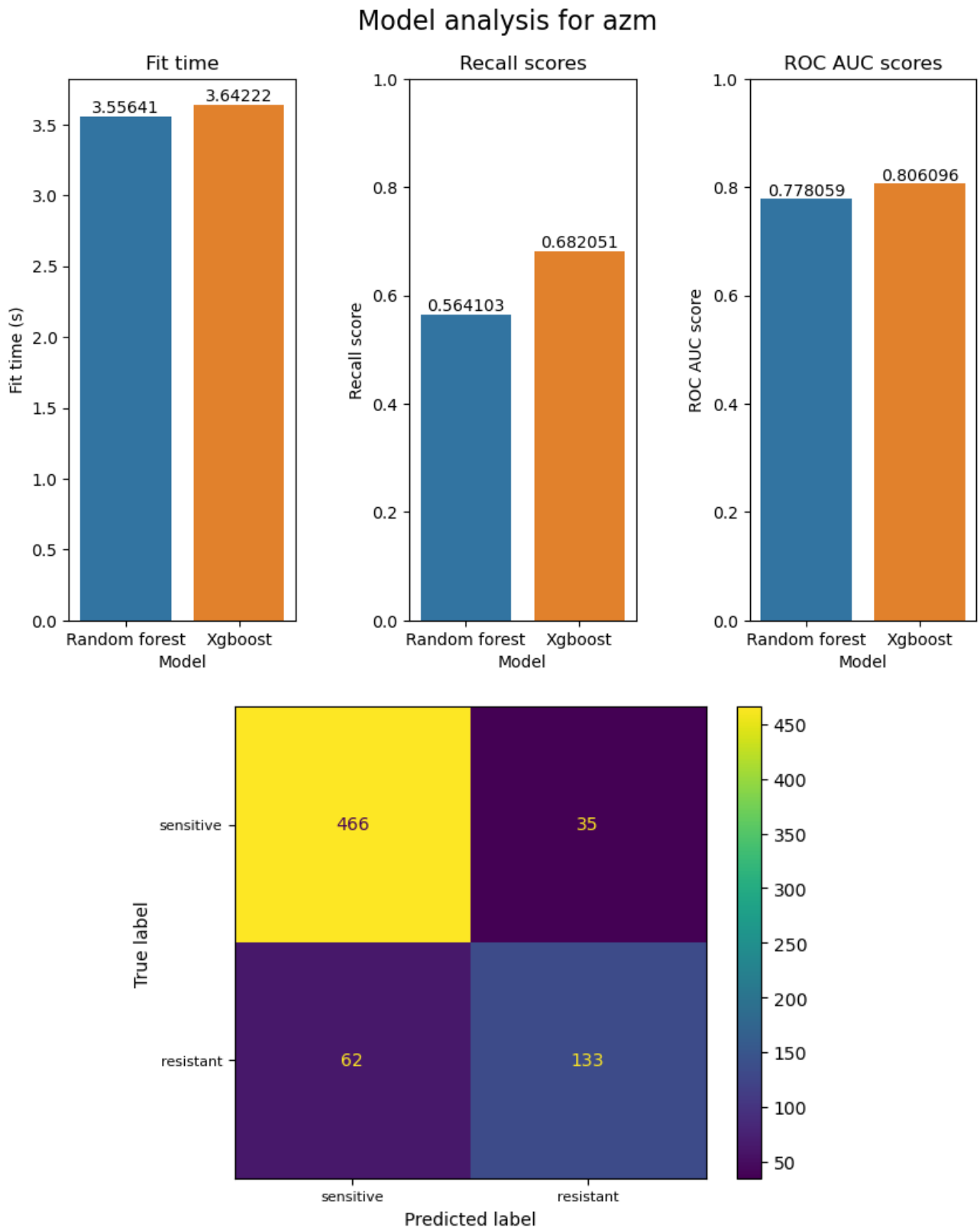


Overall, data analysis reveal unitig correlation with promise to be used in separation but without clear factors to separate and rely on for all resistant strains.

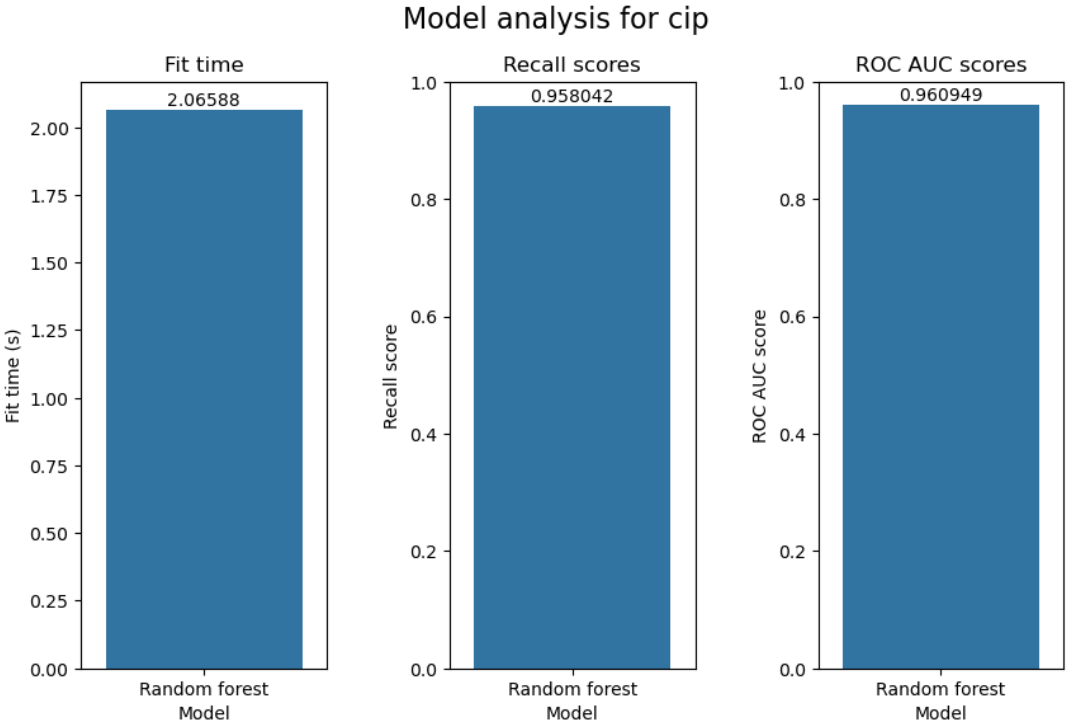
Model Selection

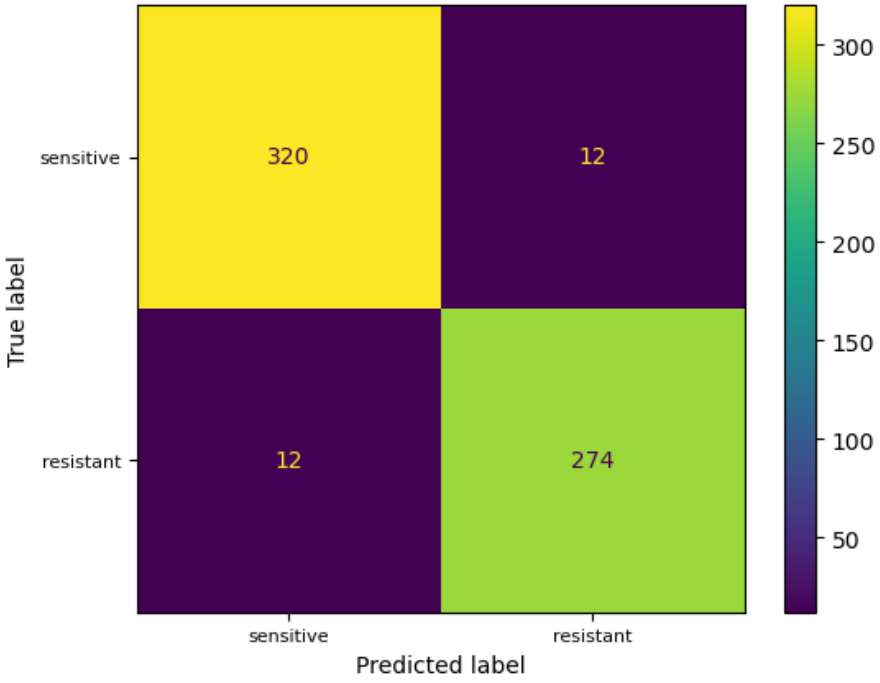
While modeling, I tested train/test splits of 80/20 and 70/30 on random forest and Xgboosting models, which are suitable for this entirely binary dataset. I optimized the models for recall, as false negatives would lead to physicians prescribing an ineffective antibiotic. Due to the large amount of features, I also took fit time into account when selecting the best models. For each random forest model and Xgboosting model, I conducted random searches to tune hyperparameters on both train/test split.

As expected given azm's low correlation between unitigs and resistance, azm's models had the worst performance. Using Xgboosting on the 80/20 split, I was able to achieve a recall of 0.68 and an ROC AUC of 0.81. Confusion matrix revealed 62 false positives compared to the 133 true positives, so further models should be explored.

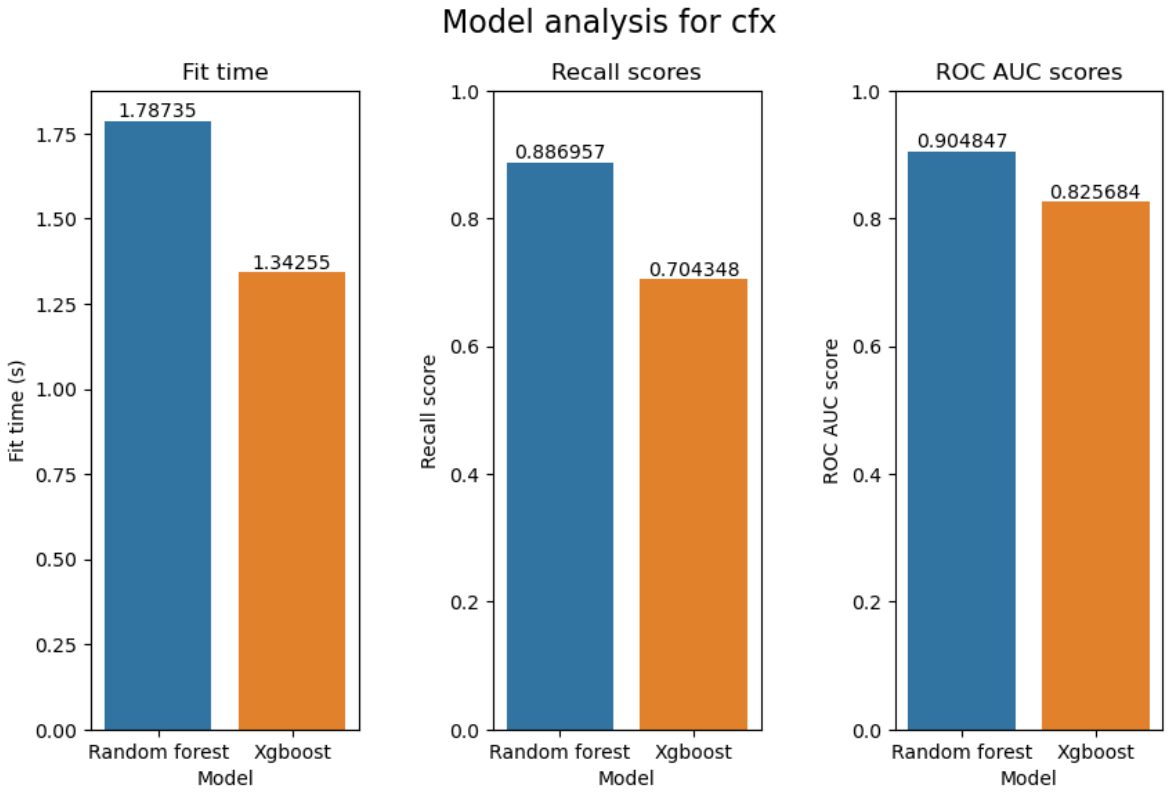

Model analysis for azm
Fit time
Fit time (s)
3.55641
3.64222
Random forest
Xgboost
Model
Recall scores
Recall score
0.564103
0.682051
ROC AUC scores
ROC AUC score
0.778059
0.806096
sensitive
resistant
True label
466
35
62
133
Predicted label

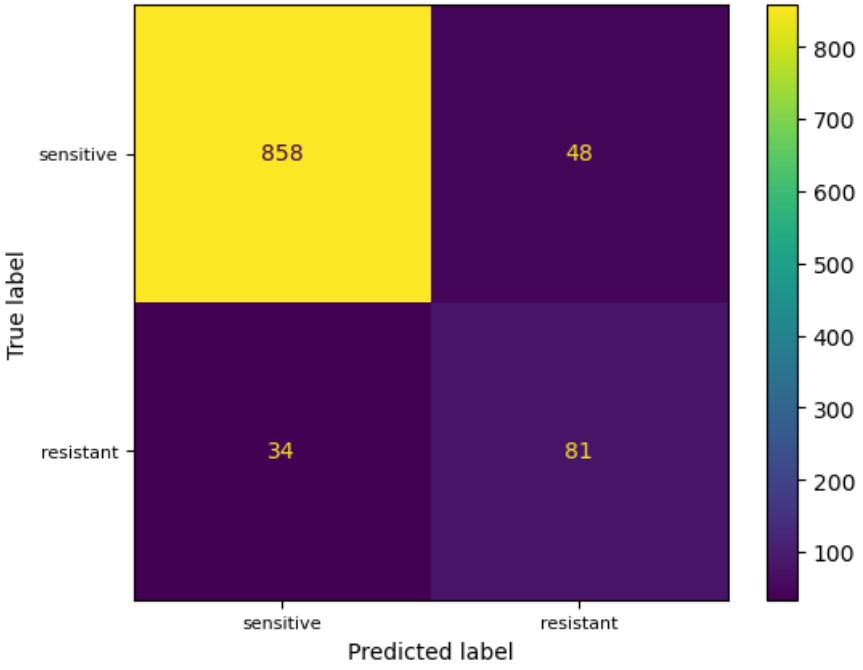
Cip's models performed the best. A random forest model with the 80/20 split received a recall and AUC ROC of 0.96. Due to long fit times, Xgboosting was not further investigated for this dataset. A confusion matrix revealed only 12 false positives and false negatives, yielding the highest performance of any of our models. This can be attributed to the higher class balance as well as the more correlated unitigs with resistance due to the increased presence of cip resistance in strains.



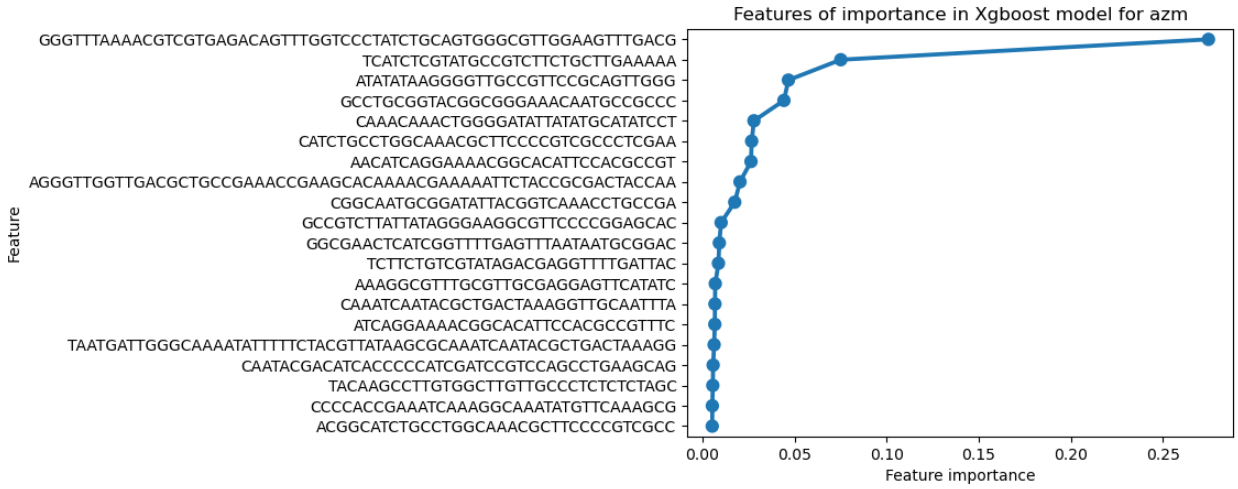


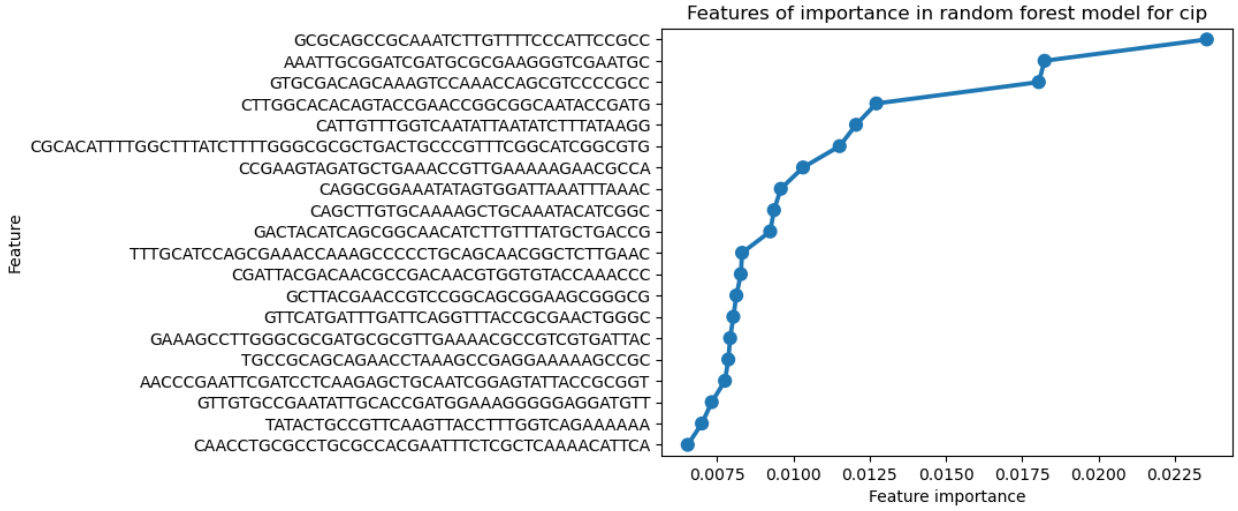
Lastly, cfx models performed surprisingly well given the class imbalance. The random forest model on the 70/30 split attained a recall of 0.89 and an ROC of 0.90. However, it did still misclassify 34 resistant samples, leading to a suggestion for further optimization.

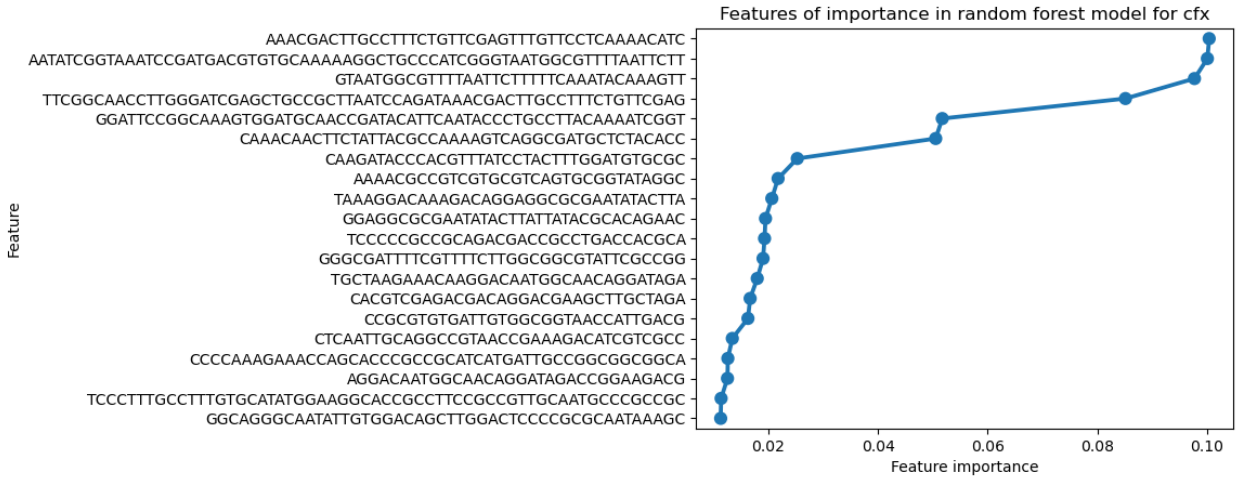




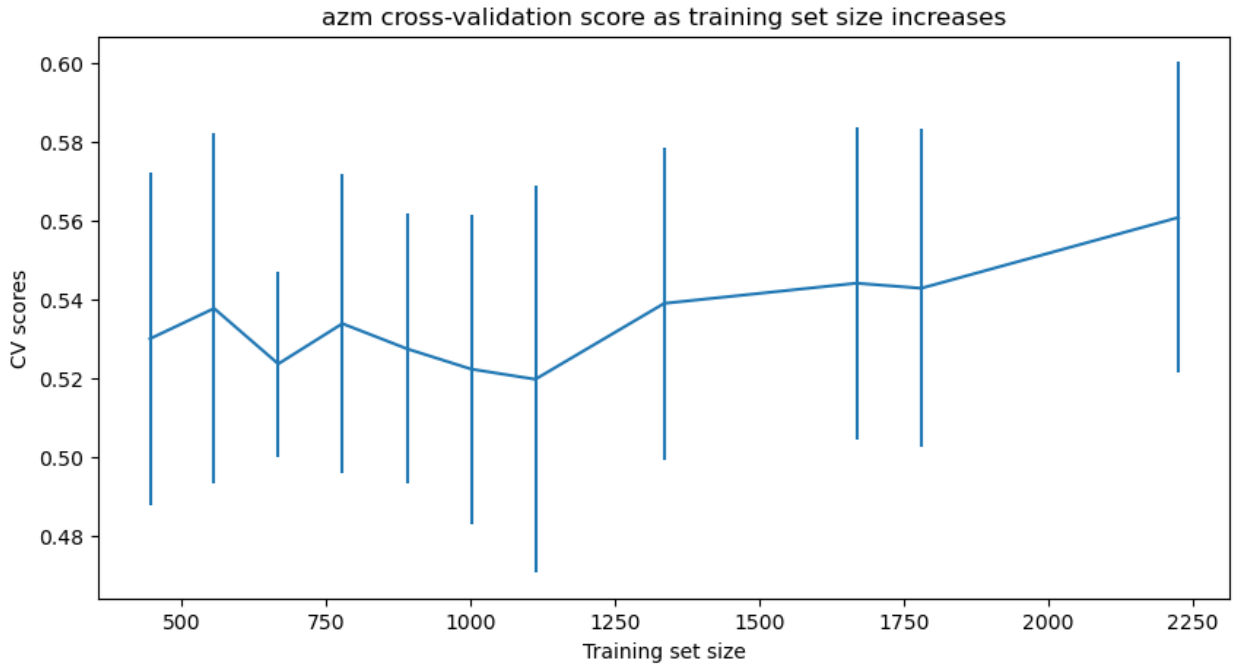
When investigating feature importance of the three models, the most important features aligned with the highly correlated unitigs discovered in data analysis. Additionally, azm relied heavily on the one unitig noted in my analysis while cip and cfx both had multiple features of high importance, indicating that azm may require another approach.

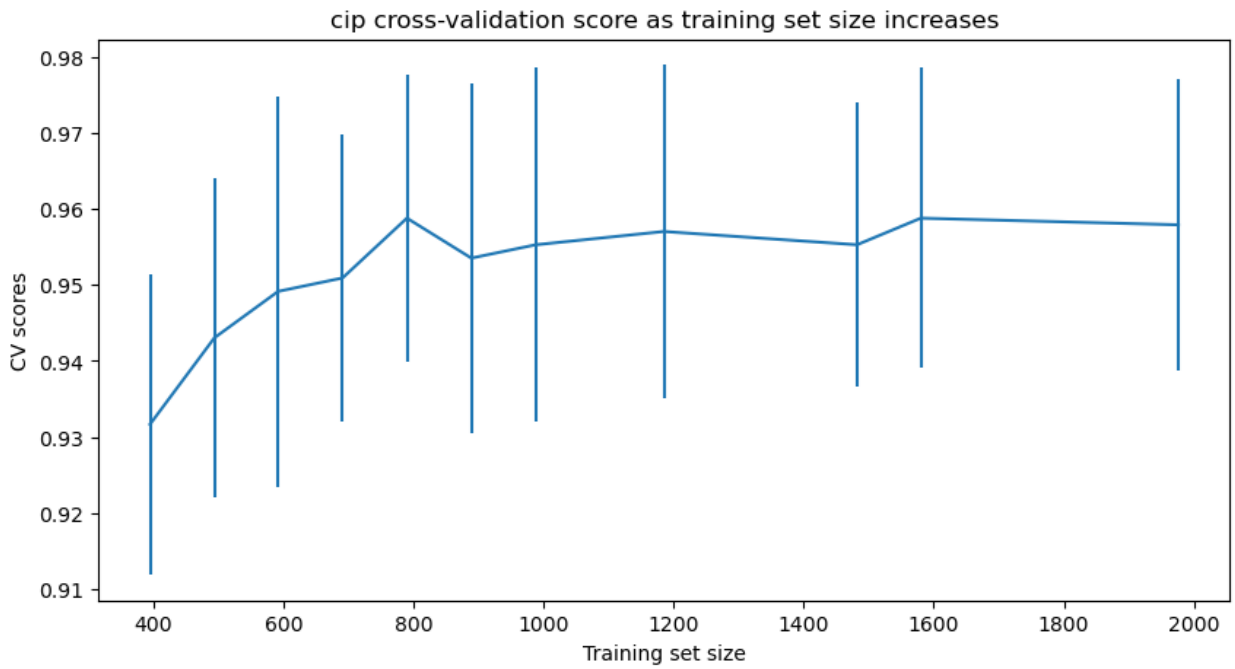


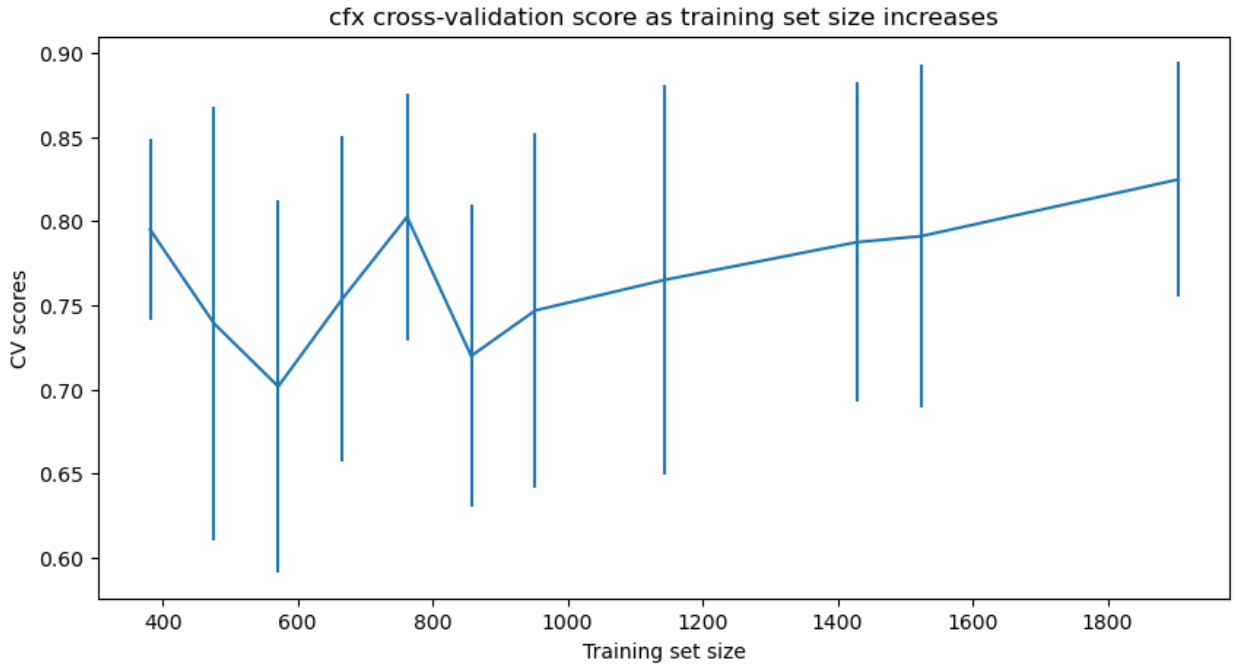




Furthermore, learning curves indicate that increasing the training sizes may further increase the performance of the models.





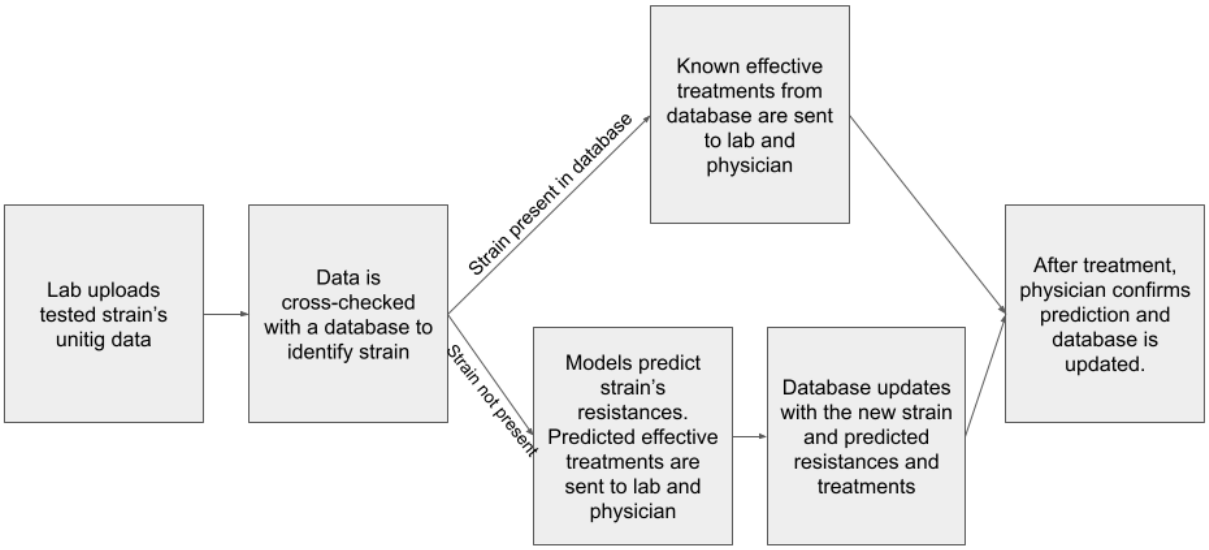


Conclusions and Future Directions

Overall, I created three models with decent to excellent recall to predict a gonorrhea strain's resistance to three different antibiotics. Azithromycin poses the greatest area for improvement within these three models. Future experiments with other models such as AdaBoost or neural nets may present opportunities to provide more

accurate predictions. I also recommend gathering more unitig data for gonorrhea strains, which will help performance of all models but especially azm as more patterns come to light. Additionally, as strains mutate frequently, updated data is crucial for the performance of such models. Another area for exploration is performing further feature selection using the features of high importance in these models. This will massively cut down the size of the datasets and improve efficiency.

Future projects include creating more models for other antibiotics, expanding into other strains of bacterial infections, and the implementation of these models to recommend treatments to physicians. This would include building a database of gonorrhea strains and their unitigs for quick assessment and data gathering.



Overall, this flow will build a database that allows for quick analysis of DNA data from gonorrhea strains and informs physicians of known effective antibiotics and predicted effective antibiotics. The gathering of this data can also allow for further training of the model once physicians have confirmed whether or not the predicted treatments are effective.