d =[d =[

Replicating and Improving the Prediction of Antidepressant Response Using the CAN-BIND and STAR*D Datasets

John-Jose Nunez
68204080
Depts. of Psychiatry and Computer Science
University of British Columbia

Teyden Nguyen
10542595
Depts. of Statistics and Computer Science
University of British Columbia

Yihan Zhou
99244733
Dept. of Computer Science
University of British Columbia

1 Introduction

Major depressive disorder (MDD) is the second-leading cause of disability globally [7]. This depressive disorder is marked by low mood, anhedonia (inability to experience joy) and neuro-vegetative symptoms such as low energy, poor concentration, and over- or under-sleeping [1]. The condition can affects all aspects of life, from hampering or even preventing work [6], to increasing mortality from other medical conditions [26]. Understandably, depression is a common reason to access healthcare, and is associated with significant healthcare costs [23].

The mainstay and first-line therapy physicians use for MDD remains prescribing an antidepressant, a variety of medications of different mechanisms [12]. While effective, only a little over half of patients will respond to this initial therapy [8], and many will need to try a different agent. Many options are available for patients who do not respond to initial treatment with an antidepressant, though it can takes months of trial and error until these options are attempted [12].

This presents a promising application for predictive modelling. If a patient can be predicted to have a high chance of not responding to treatment with an antidepressant, treatment outcomes may be improved by giving these patients more aggressive therapy sooner. For example, a high-risk patient might benefit from having a second, adjunctive medication right away, or may be prioritized to also engage in psychotherapy (counselling). Both have been shown to increase depression response rates [12], but have barriers to be used widely such as costs or side-effects.

1.1 Related Work

A recent systematic review found 26 published studies using machine learning to predict outcomes in depression [15]. Of these, 13 studies used neuroimaging data (CT and MRI scans of the brain) while three used genetic data such as DNA and micro-RNA. The authors found 7 studies using only clinical data. This type of data includes demographics (eg. income, age, education) and quantitative measures of a patient's baseline symptoms such as the severity of depressive symptoms. The authors missed the most recent study using only clinical data, by Nie et al, published in 2018 [19].

It is ideal to make predictions using only clinical data as this is easier to use clinically. Most patients with depression do not get expensive neuroimaging or genetic tests. If a physician were to use a prediction in the office, they would need it to be based on relevant clinical data. As such, this study

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

focuses on this application, and the prior work in this vein. A variety of approaches have been used to date, with various goals and methods. All have had to consider feature selection, as the categorical nature of many clinical data (eg. current medication) quickly leads to a large numbers of features relative to the number of examples. Two studies focused on using unsupervised methods to find which data are most predictive. Jain et al. [10] used hierarchical clustering and a feature search algorithm based on receiver-operator curve (ROC) performance, while Thibodeau et al. [25] employed latent-factor analysis with growth matrix modelling.

Of all work focusing on supervised methods, only Serretti et al. [22], by far the oldest study, used neural networks - a network involving one hidden layer with thirteen interconnected nodes. More recently, groups have used supervised methods based on decision trees and linear classifiers, usually employing regularization for feature selection. Iniesta et al. [9] used elastic-net with 41 features being used. Kautzky et al. [11]'s work featured the use of random forests on a 47 feature data-set. Chekroud et al. [3] use gradient boosting machines (GBM) on features chosen using elastic net, and then go on in their following paper to try different feature selection methods using unsupervised clustering [4]. The most recent work, by Nie et al. [19] compare various predictive and feature selection methods. They test logistic regression with L2 and L1/L2 regularization (elastic-net), random forests, gradient boosted decision trees, using elastic net for feature selection, as well as a method termed "clustering-χ^2" which uses K-means clustering to pick features closest to centroids, and then picks the top features based on a χ^2 score.

1.2 Contribution

Our projects seeks to continue this work improving the prediction of antidepressant response from clinical data. We do this in three ways: creating a reproducible, transparent, and reusable automated method for data processing, replicating the recent work by Nie et al. [19] and externally validating their model on a new dataset, and by testing new, previously unpublished methods for both prediction and feature selection.

Perhaps the greatest contribution to the field of applied machine learning, especially in medicine/psychiatry, will be our work on data processing. The studies to date have all used data from a handful of large psychiatry studies such as Sequenced Treatment Alternatives to Relieve Depression (STAR*D) [21]. This data is older, contains many mistakes, and contains a lot of categorical data, usually resulting in over 500 features when processed. There are many decisions that must be made for the data processing, such as how to encode various data types, and how to deal with missing data. However, there is no standardized, transparent way to do this processing. When we asked members of the team behind Nie et al. [19], they were unable to provide their processing instructions or data due to concerns about privacy, and because they processed the data manually. This hampers the field considerably; reproducibility is limited as processing may be done differently, and those wishing to do this work must devote considerable amount of time to processing before they have workable data.

Our project seeks to rectify this by conducting data processing entirely by automated, reusable, extensible software which will be publicly available upon publishing. This will allow groups to both save time, and work with the exact same data as our study, once the raw data is obtained through the National Institutes of Mental Health (NIMH). Neither the raw data nor the processed data may be shared publicly due to the privacy concerns of clinical data. Additionally, we believe our data processing may be better than prior attempts due to improved domain knowledge, as our group contains a physician practicing in psychiatry, which the group behind Nie et al. [19] did not include.

Replicability is a foundation of the scientific method. It is especially important for this field, when data processing can be done in such different ways, and prediction and feature selection methods have so many degrees of freedom. Therefore, we replicate Nie et al. [19]'s recent high-impact work, and also externally validate their model on the CAN-BIND-1 dataset [14]. This external validation is important, as to date, many of these studies have only used cross-validation to judge accuracy. CAN-BIND-1 is a very recent psychiatric depression study whose main results have yet be published. This modern data will improve the real-world applicability of these models, as the prior datasets are often over ten years old.

Lastly, our project seeks to contribute to the field by trying different methods for both feature selection and prediction. Of note, only one published study has used a neural network for this task, and there are none with more than one hidden layer [15]. Publishing results of using a deep neural network,

even if not more effective than other methods, will contribute by establishing a baseline. We will also try other methods, such as an ensemble method of prior work, a strategy we also cannot find in the extant literature. We will also explore visualization with dimensionality reduction, another strategy that seems uncommon in prior work.

To aid grading, we will briefly outline the immense amount of work spent on this project, especially for the data processing. Over 3000 columns of data between the STAR*D and CAN-BIND datasets were manually evaluated for relevance and usability, many of which needed unique processing instructions. This involved over 2500 lines of code and over 120 person-hours for the data processing itself, in addition to dozens for the actual machine learning.

2 Methods

2.1 Data and Data Processing

The clinical data used for the project comes from the STAR*D ($n = 4046$) and CAN-BIND-1 ($n = 324$). Both trials tested different treatment strategies for patients with MDD, involved patients who gave informed consent, and were overseen by clinical ethics boards. Data collected includes demographics (eg. income, persons in household, education, marital status), measures of function (eg. hours missed from work due to illness) and burden of depression or anxiety symptoms (eg. sleep quality, suicidality, energy). Patients or their treating clinicians produced the data by filling out verified scales such as the Montgomery-Asberg Depression Rating Scale [18] or the Lam Employment and Productivity Scale [13]. Data from STAR*D was obtain from the NIMH, and for CAN-BIND-1 was obtained via internal UBC servers from Dr. Raymond Lam, who is part of this study.

Traditional approaches for data cleaning, collating and engineering often use one-time-use pipelines or may often be done through a spreadsheet software such as Excel. This approach may work for the start of a project but often, as the ML pipeline is run and data requirements change, the changes to the data cleaning pipeline can start introducing errors and decrease reproducibility in research.

Therefore, we instead sought to build our pipeline similar to production level software. We built it to be highly reproducible by employing standardized libraries, using modular design, and using global configuration files. The following features of our code demonstrate these aspects:

1. Using pandas [17], a widely used data handling Python library using well tested code to reduce the need to create basic functions. We also leveraging its powerful data manipulation features needed for machine learning, such as its method for one-hot encoding.
2. As much as possible, decoupling processes such as data aggregation and data cleaning, and avoiding the use of one-time functions
3. Creating functions that can be used at many parts of the pipeline so as to minimize repeated code and reduce changes to code
4. Employing configuration files that handle common use cases for on-the-fly changes to data cleaning requirements
5. Using version control to keep track of changes over time
6. A future plan and capacity to incorporate unit and integration tests that will drastically reduce bugs and ensure data integrity, a crucial step that can not be done with Excel — this was not implemented due to time constraints

Using our pipeline was straightforward with the CAN-BIND data, which is relatively clean, as it was designed in part to be used for machine-learning. It was more difficult for STAR*D due to the poor data quality, with many labels missing, such as the week when data was recorded. There were also multiple data type inconsistencies within features. This meant multiple iterations of our pipeline for STAR*D. It also meant we did not include a handful of data likely included by Nie et al. [19] due too much ambiguity; we plan to incorporate these with more time and consultation with that group.

In order to externally validate our models, we prepared datasets from both STAR*D and CAN-BIND featuring only overlapping features. Some features were converted, such as the different depression scales used.

2.2 Feature Selection and Prediction: Replication and External Validation

Feature selection: We replicated all models used by Nie et al. [19] including the various methods for prediction and feature selection. We followed their instructions for both feature selection methods. In the first, we used elastic net to find the thirty features with largest weights. In the other, we transposed the data matrix, and then applied k-means clustering. We then picked the feature closest to the centroid as representatives of the whole cluster (vector quantisization). We then calculated χ^2 scores and selected the top n features based on this.

Prediction: We implemented the 30 models as closely as possible to Nie et al. [19]. As in that paper, we used scikit-learn [20] for the logistic regression, gradient boosted decision trees (GDBT), and random forest models. We implemented all methods from scratch, but when possible, used hyperparameters from the replicated paper, whose code we had limited access to. XGBoost [5] was downloaded from their website. We calculated accuracy as well as the area-under-curve (AUC) for the receiver-operator-curve (ROC) on a testing dataset comprised of 20% of the available data. As the replicated paper did not mention how elastic-net was implemented, we used scikit-learn's SGD model with logistic loss and L1/L2 regularization.

We externally validated the models by training them on a dataset from STAR*D comprised of overlapping features, and then tested these predictions on a CAN-BIND dataset also featuring only overlapping features.

2.3 Feature Selection and Prediction: New Methods

Agglomerative feature selection: Noticing that Nie et al. [19]'s method using χ^2 and k-means was unstable and relied on the uncommon practice of transforming a data matrix, we set out to develop a second feature selection method, based on work by Bühlmann et al. [2]. This method adopts a greedy strategy to cluster features. Initially each feature is a cluster, then the algorithms computes the given linkage distance and merges two clusters with minimal distance recursively. This feature clustering algorithm is implemented in scikit-learn [20]. We add the label value as the first feature and select the features in the same cluster as the label value because it is a reasonable assumption that features in the same cluster are closely related to each other. This feature selection method is fast but it has two drawbacks. The first drawback is that some features may be correlated to the label but have a far distance from the label, so will not be clustered together. A potential fix is to select some features from other clusters additionally so the set of selected features will be more representative; we may try this fix in the future. The other drawback is that we can only set the number of clusters, but do not know how many features will be selected. It is also difficult to measure the correlation of features within a cluster.

Prediction: We tried three different methods not used by Nie et al. [19] - support vector machines (SVMs), neural networks, and an ensemble method. The latter two have not been used before in published work in this application. Scikit-learn was again used for implementing SVMs and the neural network. We tested different structures of neural networks. The network with one hidden layer and 50 neurons has the best performance among tested models. We found ReLu was the best performing activation function, and used softmax loss as this is a binary classification problem. Weight-decay is added to the model to mitigate overfitting. We use stochastic gradient descent as the optimizer to train the neural network. Improvement is likely possible, as there are many combination of hyperparameters, optimizers and regularizers still untested. To try out an ensemble method, we made "Ultra-Ensemble", an ensemble of methods used previously: random forests, regularized logistic regression, GBDT, SVMs, and neural networks. For each method, 10 models were generating using boostrapping, and the ultimate prediction was based on the mean prediction.

Visualization: We used principal component analysis (PCA), t-SNE [16] and ISOMAP [24] to reduce the dimensionality and allow 2-D visualization. We did this to determine if any of these methods would cluster responders separately from non-responders.

3 Results

3.1 Replication and External Validation

Generally our replication of Nie et al. [19]'s methods proved comparable, but with somewhat less accurate results as in Table 1. Our AUC's were usually around 0.05 to 0.10 less. The best performing model for both groups was GDBT, though our elastic-net was our worst model, compared to theirs being l_2-penalized logistic regression.

Our external validation, using the CAN-BIND dataset instead of [19]'s RIS-INT-93, continued this pattern but with less difference 4. Our models' AUCs are within at most 0.08 worse, and our l_2 penalized logistic regression's AUC was actually slight larger.

3.2 Trying New Methods

Of the new predictive methods tested, SVMs and the neural network did not produce improved results, as in Table 2. The SVMs likely overfit. However, our "Ultra-Ensemble" seems quite promising, and was our best performing model even compared to the replicated ones. However, when its performance was externally validated with the CAN-BIND dataset, it was more similar to others, perhaps indicating some overfitting. Alternatively, it may have taken advantage of features that were not in the overlapping datasets.

Our new method for feature selection improved performance in some models, as seen in Table 3, or in more detail in appendix Table 5. This especially made a difference in some particularly poorly performing models such as our elastic-net. This supports that improved feature selection may still lead to performance gains.

3.3 Data Visualization

Unfortunately, all three methods for 2-D visualization did not provide convincing clustering, at least from we can see. They can be viewed in the appendix Section 5.2.

Table 1: Replication of Nie et al's methods using the STAR*D dataset.

Model	Nie Accuracy	Our Accuracy	Nie AUC	Our AUC
Random Forest				
Full set of features	0.70	0.605	0.78	0.732
Top n features(n=30) by clustering	0.67	0.637	0.77	0.731
Top n features(n=31) by ELNET	0.70	0.627	0.76	0.609
GBDT				
Full set of features	0.70	0.665	0.78	0.746
Top n features(n=30) by clustering	0.70	0.688	0.77	0.780
Top n features(n=31) by ELNET	0.70	0.600	0.76	0.645
XGBoost				
Full set of features	0.67	0.598	0.76	0.675
Top n features(n=30) by clustering	0.66	0.631	0.73	0.688
Top n features(n=31) by ELNET	0.67	0.650	0.76	0.642
l_2**penalized logistic regression**				
Full set of features	0.63	0.583	0.69	0.646
Top n features(n=30) by clustering	0.71	0.641	0.73	0.699
Top n features(n=31) by ELNET	0.72	0.612	0.77	0.680
Elastic net with SGD				
Top n features(n=31) by ELNET	0.70	0.600	0.76	0.624

Nie: the results reported in [19]

4 Discussion

In our project, we contribute to the field of applied machine learning in psychiatry by replicating a recent study that used clinical data to predict antidepressant response. We externally validate this

Table 2: New methods attempted in this project using the STAR*D and then CAN-BIND datasetsl

Model	Training accuracy*	Testing accuracy	AUC
STAR*D test data set			
NN			
Full set of features	0.627	0.585	0.651
Top n features(n=30) by clustering	0.636	0.576	0.628
Top n features(n=31) by ELNET	0.634	0.598	0.634
SVM			
Full set of features	0.947	0.308	0.624
Top n features(n=30) by clustering	0.945	0.277	0.517
Top n features(n=31) by ELNET	0.655	0.627	0.654
Ultra-Ensemble			
Full set of features	0.803	0.643	0.773
Top n features(n=30) by clustering	0.786	0.594	0.765
Top n features(n=31) by ELNET	0.583	0.519	0.595
Can-Bind Dataset			
NN	0.650	0.611	0.661
SVM	0.964	0.506	0.485
Ultra-Ensemble	0.866	0.633	0.657

* Not a true training accuracy because of the subsampling method used

Table 3: New feature selection method: Agglomerative clustering with 50 clusters

Model	Training accuracy*	Testing accuracy	AUC
STAR*D test data set			
Random Forest	0.887	0.601	0.683
GBDT	0.709	0.610	0.656
XGBoost	0.639	0.609	0.670
$l_2$penalized logistic regression	0.663	0.659	0.705
Elastic net with SGD	0.662	0.656	0.707
NN	0.786	0.571	0.639
SVM	0.673	0.633	0.693
Ultra-Ensemble	0.809	0.564	0.642

Table 4: External validation of models using the CAN-BIND and RIS-INT-93 datasets

Model	Nie Accuracy*	Our Accuracy	Nie AUC	Our AUC
Random Forest	0.65	0.650	0.73	0.653
GBDT	0.64	0.622	0.71	0.675
XGBoost	0.67	0.617	0.73	0.663
$l_2$penalized logistic regression	0.64	0.578	0.60	0.614

*Nie: the balanced accuracy reported in [19] using the RIS-INT-93 dataset

study with a new, never-before-used clinical dataset. We find that the replication is imperfect, with reduced performance seen in most models. We also find that the external validation with our new dataset similarly shows worse performance. We then test new methods for both prediction and feature selection, and find mixed results, with mostly middling results, but some improvement with our ensemble method. We made an extensible, modular data processing pipeline that will be shared in the future, and found that this worked well, but was labour intensive to create.

That our replication was imperfect supports the need for increased replicability in this field, especially in data processing. Some of the reduced performance of our model may be due to differences in predictive model implementation. Generally this should be minimal, as off-the-shelf models were primarily used, such as from scikit-learn, with the same hyperparameters used in many cases. As such, a difference in data processing seems likely to be the main reason for this reduced performance. A difference seems inevitable, given the many decisions we needed to make during the data processing. This imperfect replication is both a strength and weakness of the project; better replication was hoped

for, but it is also valuable to know that replication is currently likely not possible with the given published data, therefore identifying a gap in the scientific process.

Similarly, the performance of the new models is both a strength and weakness. The relatively poor performance of our neural network and SVMs is disappointing. Their performance faces the same constraints as other models due to the data processing. However, they may be able to perform better with further tuning. Our ensemble method is encouraging, with its performance the best of all methods when used on the STAR*D dataset. It seems possible that, if our models were not penalized by the data processing, our ensemble method may indeed outperform the best model in Nie et al. [19], which represents the current state-of-the-art. This is in line with the known advantage of ensemble methods.

Our data processing pipeline will likely be a significant contribution to the field. While it took a considerable amount of time to create, it will allow others to reproduce our data and obtain processed data for both STAR*D and CAN-BIND-1 datasets quickly. This will reduce the time needed to try machine-learning methods on this data from hundreds of hours to mere minutes for future work. The standardized, modurarlzied method of coding, made possible by our group member with years of related industry experience, will allows others to easily adopt our software. This is a considerable strength of the project. However, our results suggest that our data processing was not as good as that done by Nie et al. [19]; further refinement of our data processing is likely needed; time was a constraint.

4.1 Future Work

We will be seeking to publish this project in a journal such as *Journal of Clinical Psychiatry* or a similar destination. Attesting to the amount of work done in this project, we expect little will need to be added. However, we will need to spend some time on refining our data. For instance, there are some ways to fill gaps in certain clinical data with more complicated but likely more accurate methods. This will take additional time to validate and code. We will also try conversing again with the team behind Nie et al. [19] to ask about some of the specific decisions they made in data processing, in attempt to close the performance gap. Further work on the pipeline will include refactoring the code for increased reusability, and adding a test suite that ensures data integrity between transformed data and the original.

With this project completed, a variety of new and likely high-yield investigations are possible. We may work on a separate paper for our data pipeline, using methods to validate it, and improving features to allow further extensibility. Another idea would be to use the pipeline to process more, or even all, of the similar large antidepressant studies. This could allow predictions with around an order of magnitude more data than any work so far. However, this may be difficult due to the lack of overlapping features between studies, as different studies use different clinical scales. Transfer learning is an obvious possible solution to this, and we are unaware of any study applying this to psychiatry, presenting an opportunity.

References

[1] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. en. Fifth Edition. American Psychiatric Association, May 2013. ISBN: 978-0-89042-555-8 978-0-89042-557-2. DOI: 10.1176/appi.books.9780890425596. URL: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596.

[2] Peter Bühlmann et al. "Correlated variables in regression: clustering and sparse estimation". en. In: (Sept. 2012). DOI: 10.1016/j.jspi.2013.05.019. URL: https://arxiv.org/abs/1209.5908v1.

[3] Adam Mourad Chekroud et al. "Cross-trial prediction of treatment outcome in depression: a machine learning approach". en. In: *The Lancet Psychiatry* 3.3 (Mar. 2016), pp. 243–250. ISSN: 22150366. DOI: 10.1016/S2215-0366(15)00471-X. URL: http://linkinghub.elsevier.com/retrieve/pii/S221503661500471X.

[4] Adam M. Chekroud et al. "Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach". en. In: *JAMA Psychiatry* 74.4 (Apr. 2017), pp. 370–378. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2017.0025. URL: http://jamanetwork.com/journals/jamapsychiatry/fullarticle/2604309.

[5] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[6] Alize J. Ferrari et al. "The Epidemiological Modelling of Major Depressive Disorder: Application for the Global Burden of Disease Study 2010". en. In: *PLOS ONE* 8.7 (July 2013), e69637. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0069637. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069637.

[7] Mohammad H Forouzanfar et al. "Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013". en. In: *The Lancet* 386.10010 (Dec. 2015), pp. 2287–2323. ISSN: 01406736. DOI: 10.1016/S0140-6736(15)00128-2. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673615001282.

[8] F. Hieronymus et al. "Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression". en. In: *Molecular Psychiatry* 21.4 (Apr. 2016), pp. 523–530. ISSN: 1476-5578. DOI: 10.1038/mp.2015.53. URL: https://www.nature.com/articles/mp201553.

[9] Raquel Iniesta et al. "Combining clinical variables to optimize prediction of antidepressant treatment outcomes". In: *Journal of Psychiatric Research* 78 (July 2016), pp. 94–102. ISSN: 0022-3956. DOI: 10.1016/j.jpsychires.2016.03.016. URL: http://www.sciencedirect.com/science/article/pii/S0022395616300541.

[10] Felipe A. Jain et al. "Predictive Socioeconomic and Clinical Profiles of Antidepressant Response and Remission". en. In: *Depression and Anxiety* 30.7 (2013), pp. 624–630. ISSN: 1520-6394. DOI: 10.1002/da.22045. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/da.22045.

[11] Alexander Kautzky et al. "Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample". English. In: *The Journal of Clinical Psychiatry* 79.1 (Dec. 2017), pp. –. ISSN: 0160-6689. DOI: 10.4088/JCP.16m11385. URL: http://www.psychiatrist.com/JCP/article/Pages/2018/v79n01/16m11385.aspx.

[12] Sidney H. Kennedy et al. "Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments". en. In: *The Canadian Journal of Psychiatry* 61.9 (Sept. 2016), pp. 540–560. ISSN: 0706-7437, 1497-0015. DOI: 10.1177/0706743716659417. URL: http://journals.sagepub.com/doi/10.1177/0706743716659417.

[13] R. W. Lam. "Lam employment absence and productivity scale (LEAPS): Further validation studies in major depressive disorder". English. In: *Value in Health* 17.3 (May 2014), A195. ISSN: 1098-3015, 1524-4733. DOI: 10.1016/j.jval.2014.03.1137. URL: https://www.valueinhealthjournal.com/article/S1098-3015(14)01188-7/abstract.

[14] Raymond W. Lam et al. "Discovering biomarkers for antidepressant response: protocol from the Canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort". en. In: *BMC Psychiatry* 16.1 (Dec. 2016). ISSN: 1471-244X. DOI: 10.1186/s12888-016-0785-x. URL: http://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-016-0785-x.

[15] Yena Lee et al. "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review". In: *Journal of Affective Disorders* 241 (Dec. 2018), pp. 519–532. ISSN: 0165-0327. DOI: 10.1016/j.jad.2018.08.073. URL: http://www.sciencedirect.com/science/article/pii/S0165032718304853.

[16] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of machine learning research* 9 (Nov. 2008), pp. 2579–2605.

[17] Wes McKinney. "pandas: a Foundational Python Library for Data Analysis and Statistics". en. In: (), p. 9.

[18] S. A. Montgomery and M. Asberg. "A new depression scale designed to be sensitive to change". en. In: *The British Journal of Psychiatry* 134.4 (Apr. 1979), pp. 382–389. ISSN: 0007-1250. DOI: 10.1192/bjp.134.4.382. URL: http://bjp.rcpsych.org/cgi/doi/10.1192/bjp.134.4.382.

[19] Zhi Nie et al. "Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study". en. In: *PLOS ONE* 13.6 (June 2018), e0197268. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0197268. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197268.

[20] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830. ISSN: ISSN 1533-7928. URL: http://www.jmlr.org/papers/v12/pedregosa11a.html.

[21] A. John Rush et al. "Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design". In: *Controlled Clinical Trials* 25.1 (Feb. 2004), pp. 119–142. ISSN: 0197-2456. DOI: 10.1016/S0197-2456(03)00112-0. URL: http://www.sciencedirect.com/science/article/pii/S0197245603001120.

[22] Alessandro Serretti et al. “A neural network model for combining clinical predictors of antidepressant response in mood disorders”. In: *Journal of Affective Disorders* 98.3 (Mar. 2007), pp. 239–245. ISSN: 0165-0327. DOI: 10.1016/j.jad.2006.08.008. URL: http://www.sciencedirect.com/science/article/pii/S0165032706003429.

[23] Gregory E. Simon, Michael VonKorff, and William Barlow. “Health Care Costs of Primary Care Patients With Recognized Depression”. en. In: *Archives of General Psychiatry* 52.10 (Oct. 1995), pp. 850–856. ISSN: 0003-990X. DOI: 10.1001/archpsyc.1995.03950220060012. URL: https://jamanetwork-com.ezproxy.library.ubc.ca/journals/jamapsychiatry/fullarticle/497246.

[24] J. B. Tenenbaum. “A Global Geometric Framework for Nonlinear Dimensionality Reduction”. en. In: *Science* 290.5500 (Dec. 2000), pp. 2319–2323. ISSN: 00368075, 10959203. DOI: 10.1126/science.290.5500.2319. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.290.5500.2319.

[25] Michel A. Thibodeau et al. “Latent Classes of Nonresponders, Rapid Responders, and Gradual Responders in Depressed Outpatients Receiving Antidepressant Medication and Psychotherapy”. en. In: *Depression and Anxiety* 32.3 (2015), pp. 213–220. ISSN: 1520-6394. DOI: 10.1002/da.22293. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/da.22293.

[26] Elizabeth Reisinger Walker, Robin E. McGee, and Benjamin G. Druss. “Mortality in Mental Disorders and Global Disease Burden Implications: A Systematic Review and Meta-analysis”. en. In: *JAMA Psychiatry* 72.4 (Apr. 2015), pp. 334–341. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2014.2502. URL: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2110027.

5 Appendix

5.1 Table of all models

Table 5: Model performance

Model	Accuracy on training set	Testing accuracy	AUC
STAR*D test data set			
Random Forest			
Full set of features	0.884	0.605	0.732
Top n features(n=30) by clustering(50)	0.875	0.637	0.731
Top n features(n=30) by clustering(75)	0.864	0.628	0.707
Top n features(n=30) by clustering(100)	0.870	0.605	0.716
Top n features(n=31) by ELNET	0.693	0.627	0.609
Agglomerative feature selection(50 clusters)	0.887	0.601	0.683
GBDT			
Full set of features	0.771	0.665	0.746
Top n features(n=30) by clustering(50)	0.735	0.688	0.780
Top n features(n=30) by clustering(75)	0.746	0.668	0.740
Top n features(n=30) by clustering(100)	0.732	0.684	0.777
Top n features(n=31) by ELNET	0.667	0.600	0.645
Agglomerative feature selection(50 clusters)	0.709	0.610	0.656
XGBoost			
Full set of features	0.653	0.598	0.675
Top n features(n=30) by clustering(50)	0.651	0.631	0.688
Top n features(n=30) by clustering(75)	0.654	0.585	0.655
Top n features(n=30) by clustering(100)	0.653	0.614	0.715
Top n features(n=31) by ELNET	0.662	0.650	0.642
Agglomerative feature selection(50 clusters)	0.639	0.609	0.670
$l_2$penalized logistic regression			
Full set of features	0.741	0.583	0.646
Top n features(n=30) by clustering(50)	0.661	0.624	0.706
Top n features(n=30) by clustering(75)	0.666	0.634	0.693
Top n features(n=30) by clustering(100)	0.662	0.641	0.699
Top n features(n=31) by ELNET	0.668	0.612	0.680
Agglomerative feature selection(50 clusters)	0.663	0.659	0.705
Elastic net with SGD			
Top n features(n=31) by ELNET	0.668	0.600	0.624
Agglomerative feature selection(50 clusters)	0.662	0.656	0.707
NN			
Full set of features	0.627	0.585	0.651
Top n features(n=30) by clustering(50)	0.619	0.579	0.667
Top n features(n=30) by clustering(75)	0.677	0.605	0.626
Top n features(n=30) by clustering(100)	0.636	0.576	0.628
Top n features(n=31) by ELNET	0.634	0.598	0.634
Agglomerative feature selection(50 clusters)	0.786	0.571	0.639
SVM			
Full set of features	0.947	0.308	0.624
Top n features(n=30) by clustering(50)	0.945	0.277	0.517
Top n features(n=30) by clustering(75)	0.949	0.265	0.542
Top n features(n=30) by clustering(100)	0.949	0.257	0.525
Top n features(n=31) by ELNET	0.655	0.627	0.654
Agglomerative feature selection(50 clusters)	0.673	0.633	0.693
Ultra-Ensemble			
Full set of features	0.803	0.643	0.773
Top n features(n=30) by clustering(50)	0.808	0.589	0.722
Top n features(n=30) by clustering(75)	0.786	0.594	0.765
Top n features(n=30) by clustering(100)	0.817	0.571	0.721
Top n features(n=31) by ELNET	0.583	0.519	0.595
Agglomerative feature selection(50 clusters)	0.809	0.564	0.642
Can-bind data set			
Random Forest	0.929	0.650	0.653
GBDT	0.736	0.622	0.675
Elastic net with SGD	0.662	0.594	0.642
XGBoost	0.693	0.617	0.663
$l_2$penalized logistic regression	0.656	0.578	0.614
NN	0.650	0.611	0.661
SVM	0.964	0.506	0.485
Ultra-Ensemble	0.866	0.633	0.657

5.2 Visualization

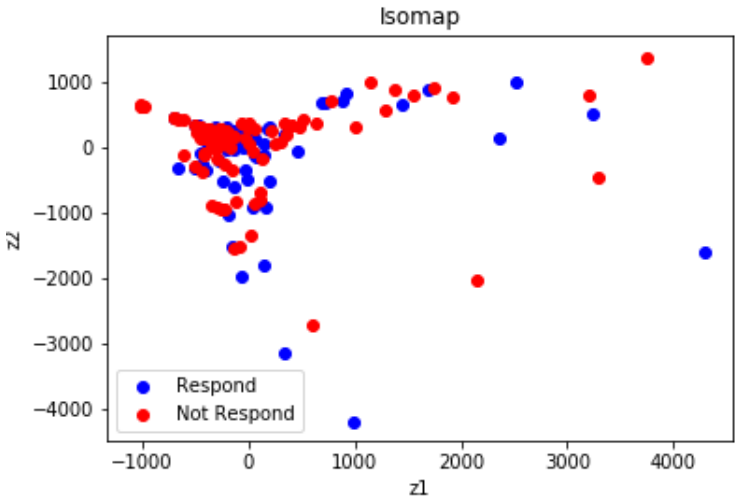

Figure 1: Isomap for Can-Bind

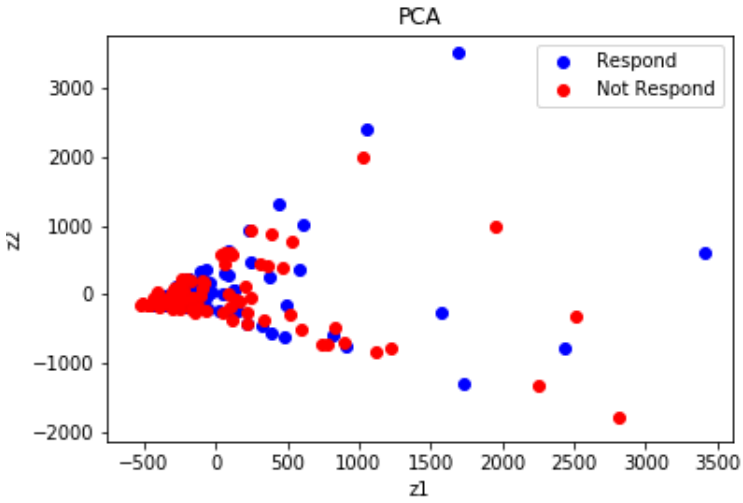

Figure 2: PCA for Can-Bind

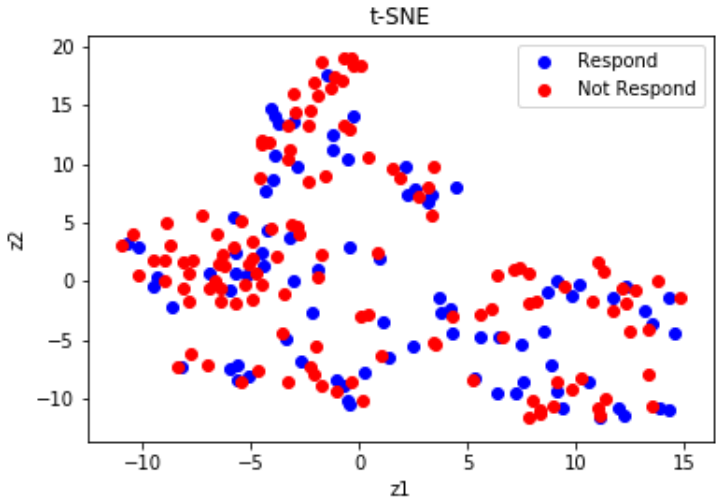

Figure 3: t-SNE for Can-Bind

5.3 Comparison between models

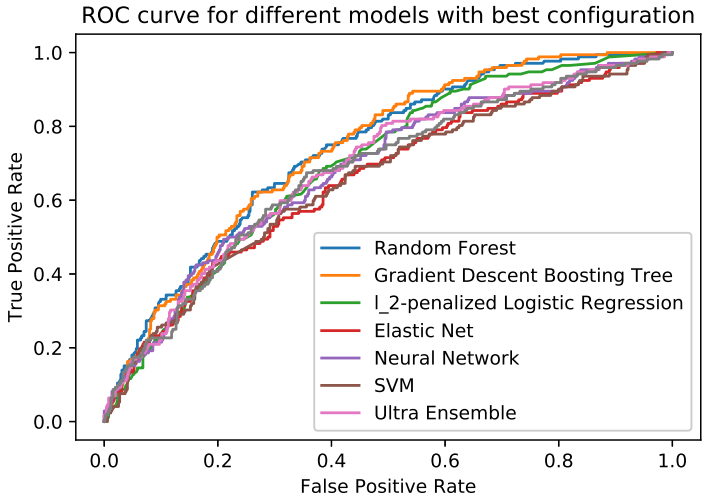

Figure 4: Comparison on Star*D

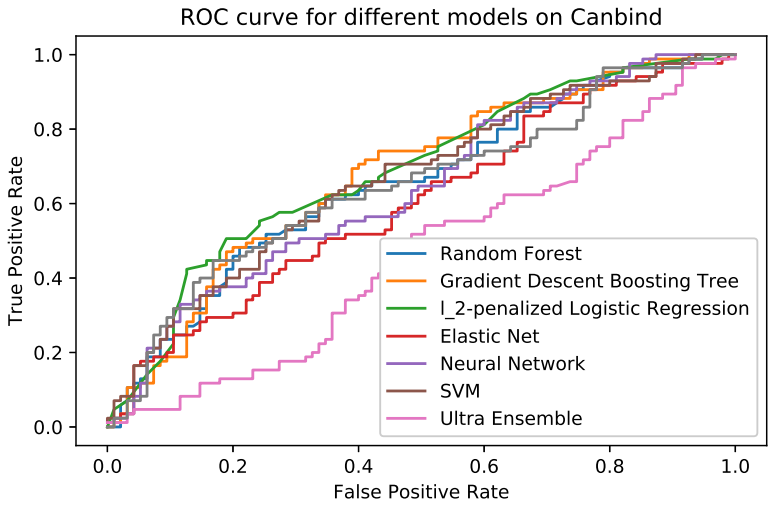

Figure 5: Comparison on Can-Bind

5.4 Receiver Operating Characteristic Curves

The following figures represent the ROCs of various methods tested. They are captioned by the acronym of the model used, with the number indicated the order as found in Table 5.

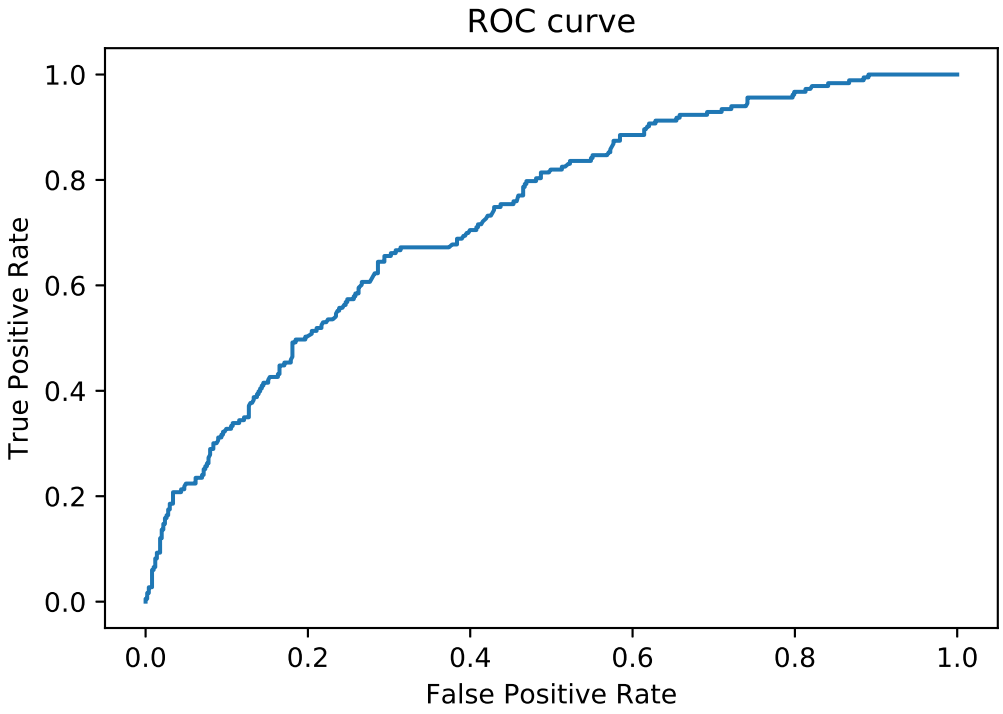

Figure 6: RF-1

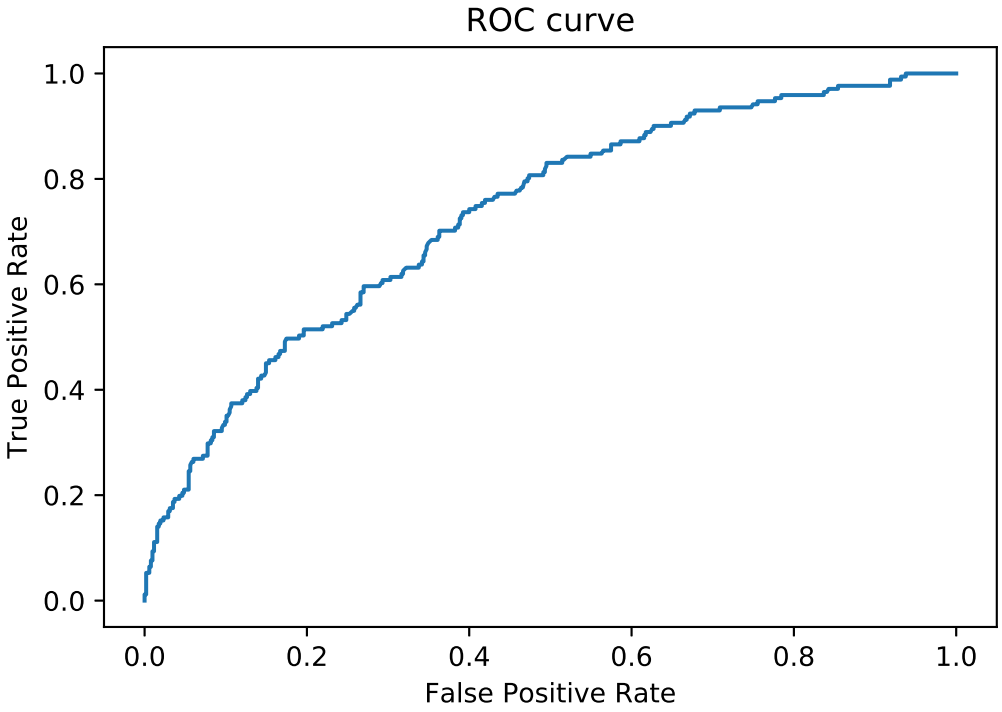

Figure 7: RF-2

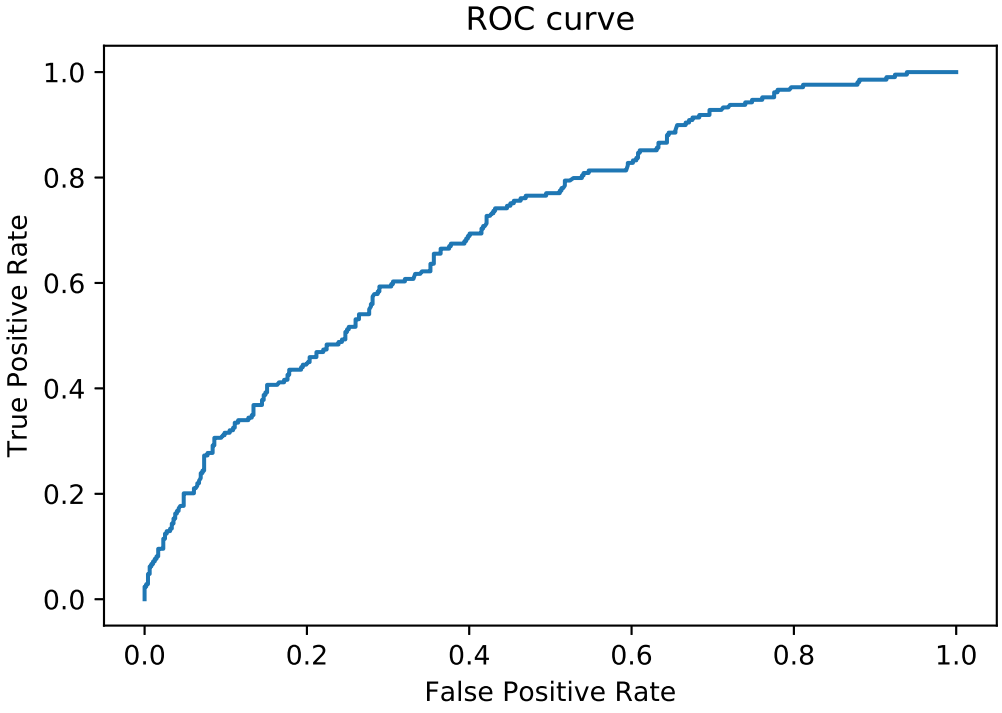

Figure 8: RF-3

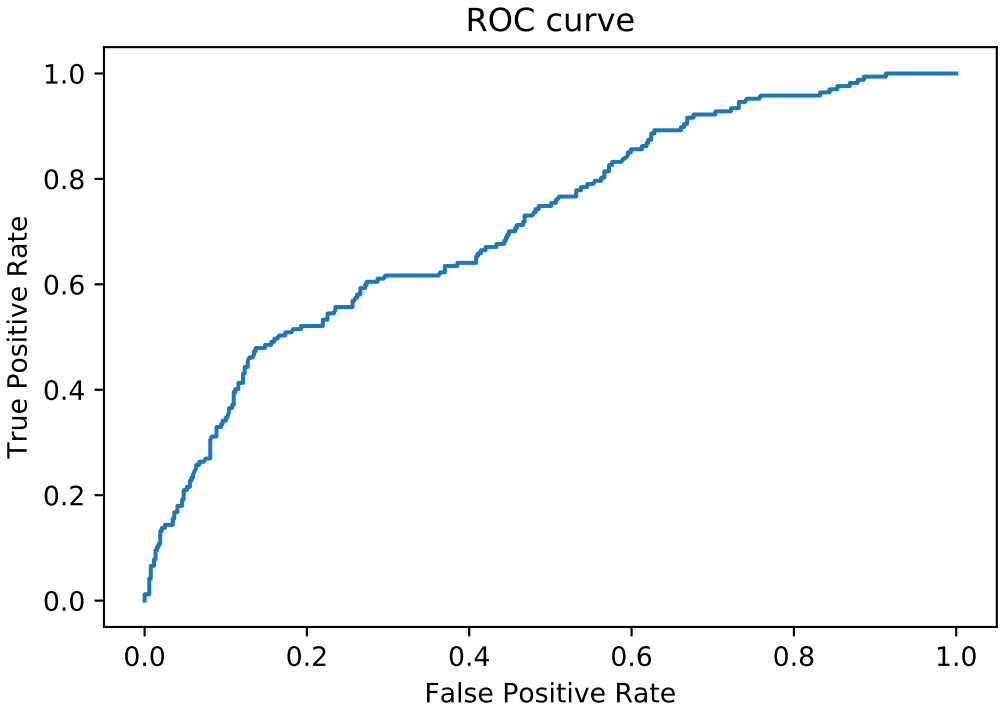

Figure 9: RF-4

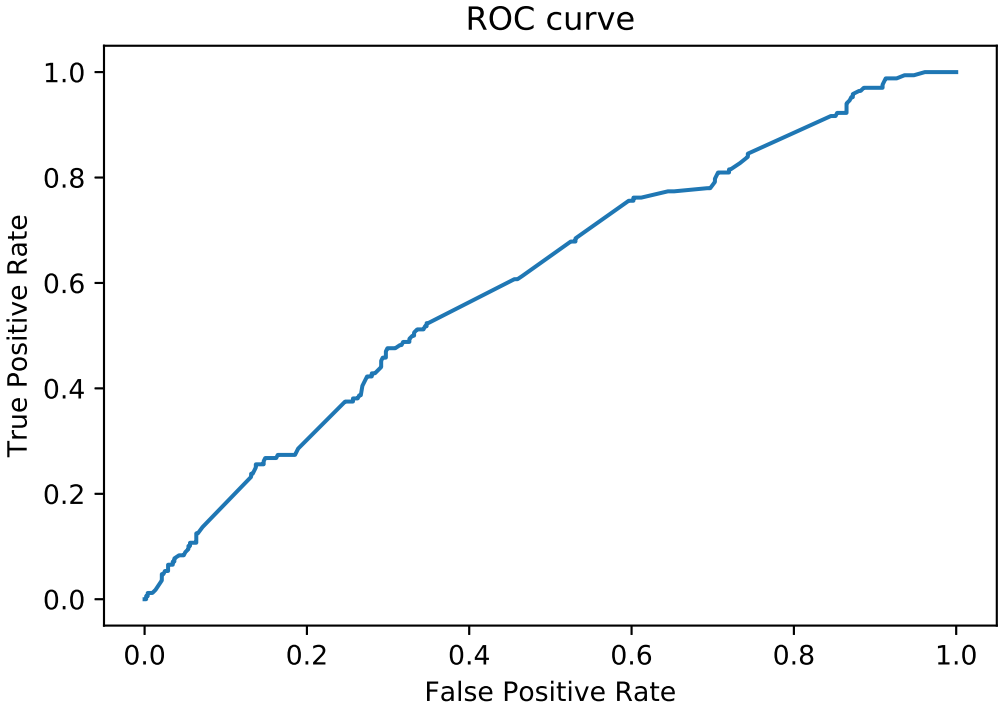

Figure 10: RF-5

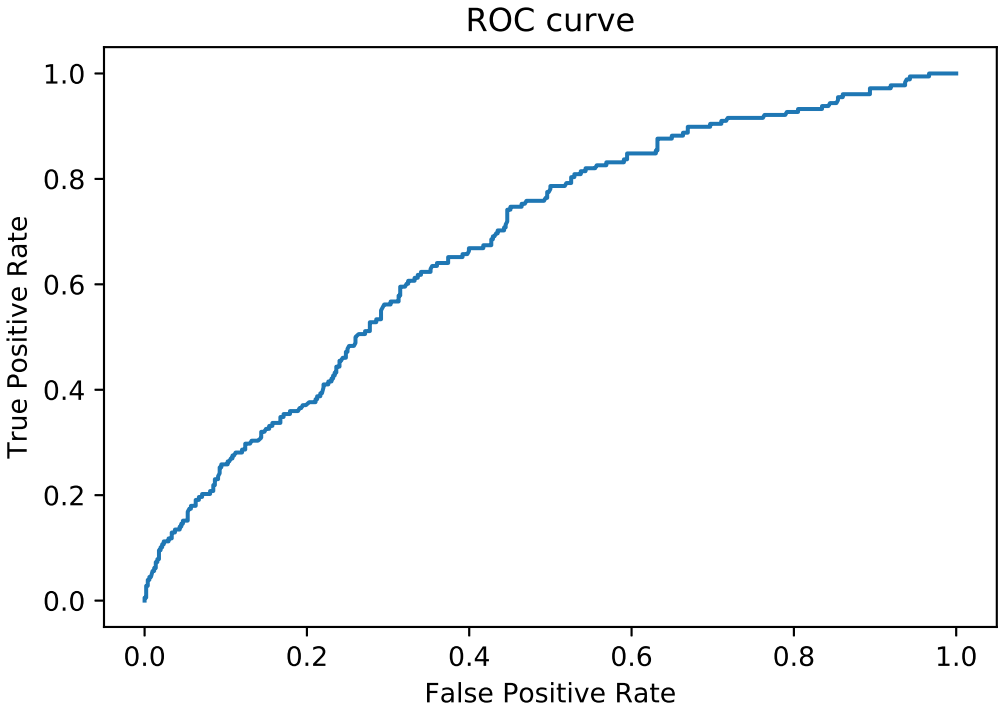

Figure 11: RF-6

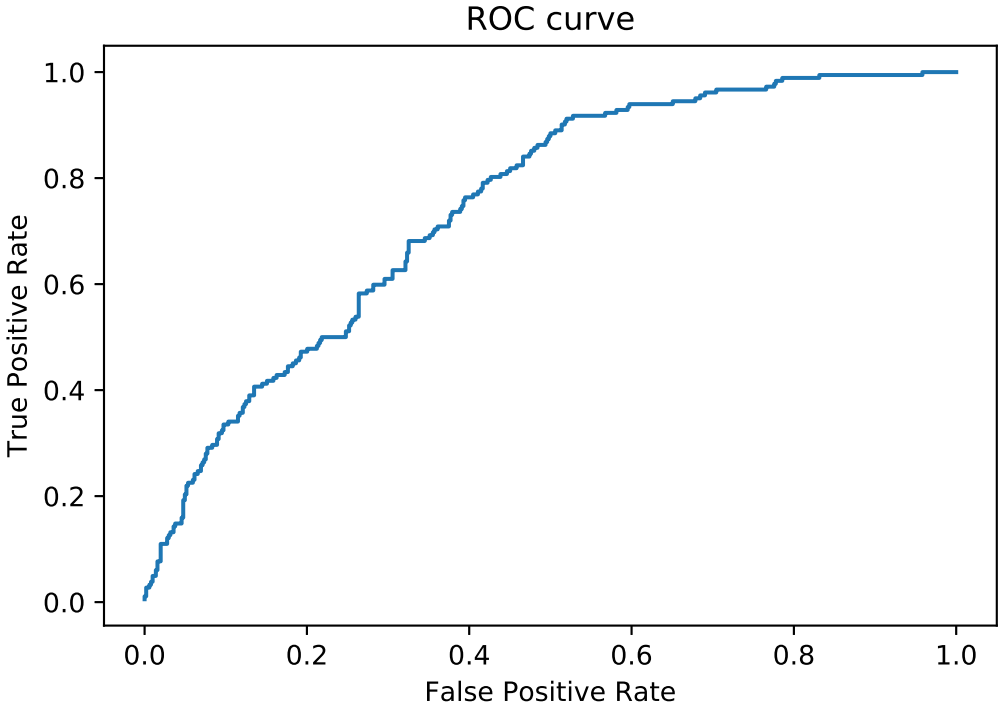

Figure 12: GBDT-1

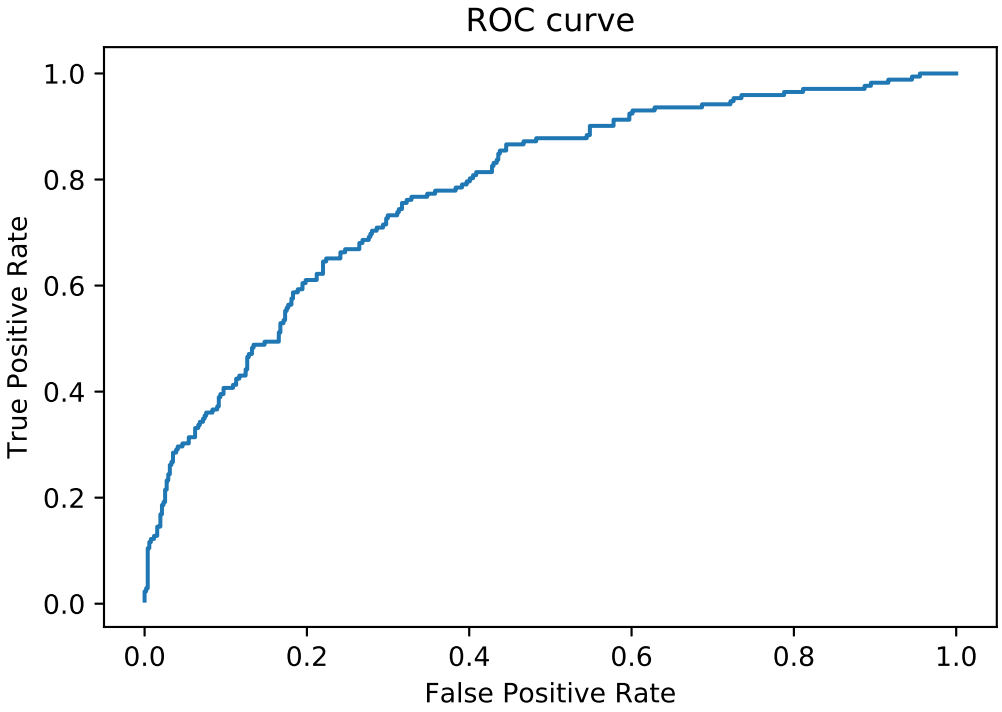

Figure 13: GBDT-2

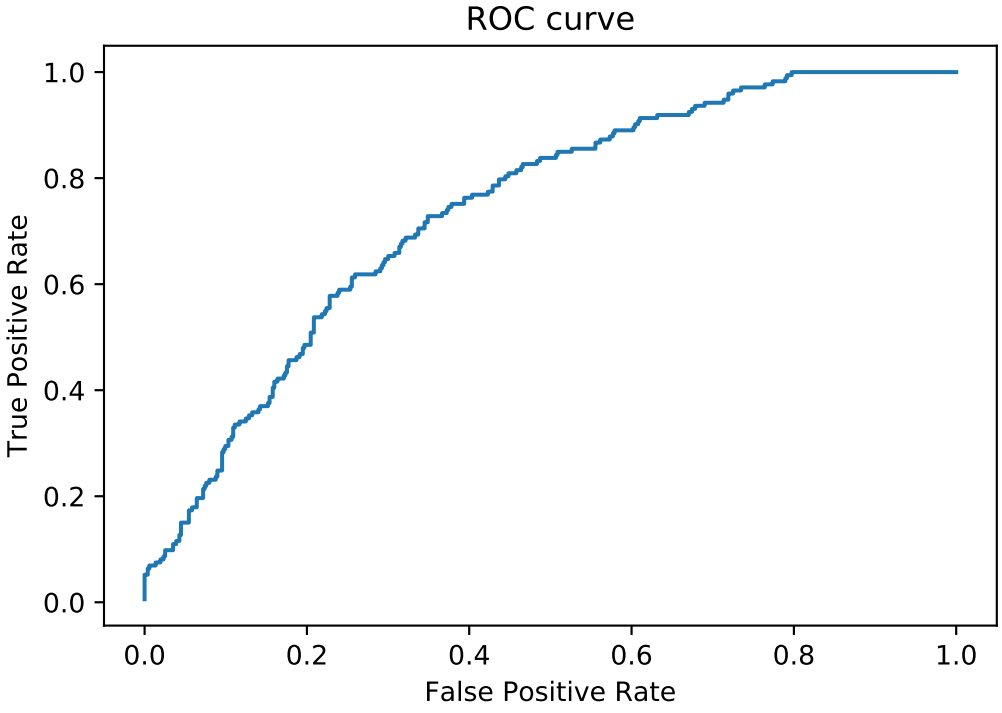

Figure 14: GBDT-3

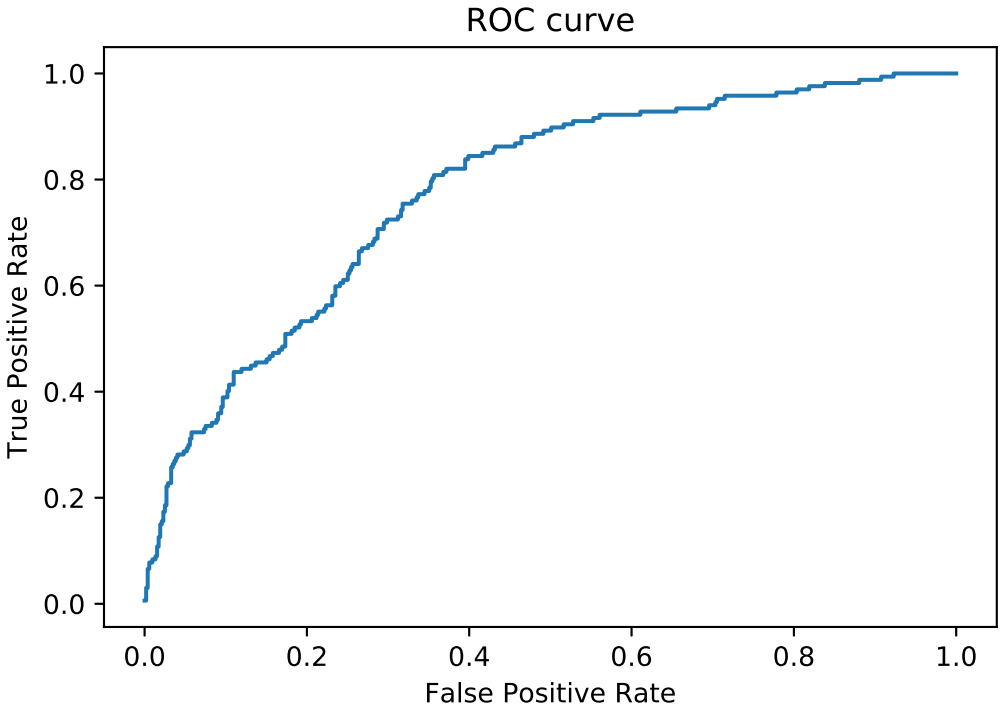

Figure 15: GBDT-4

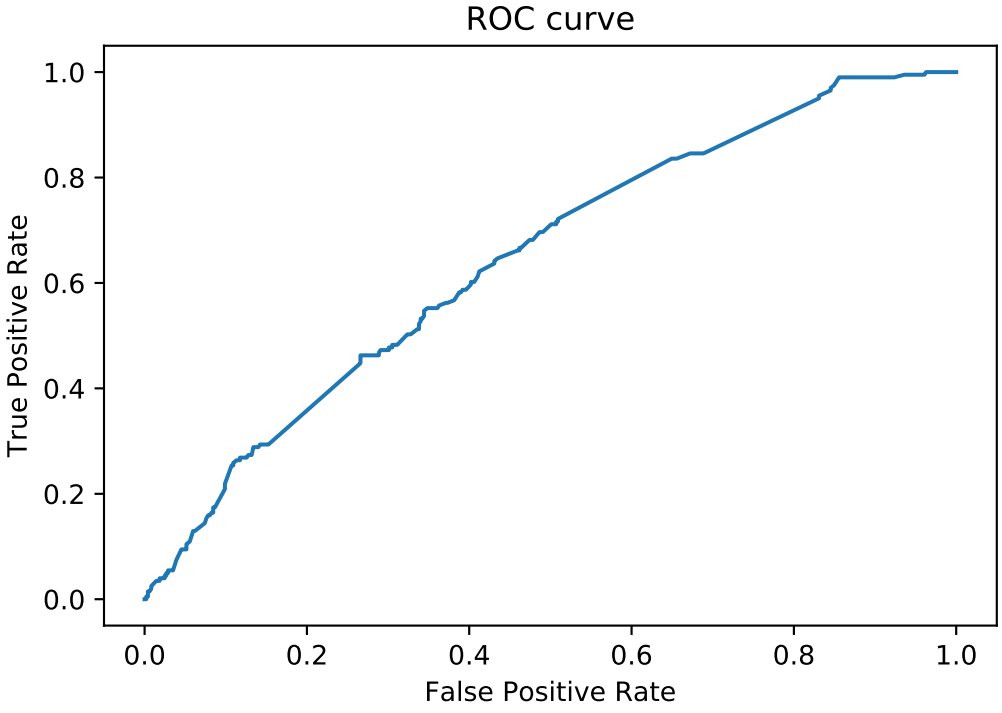

Figure 16: GBDT-5

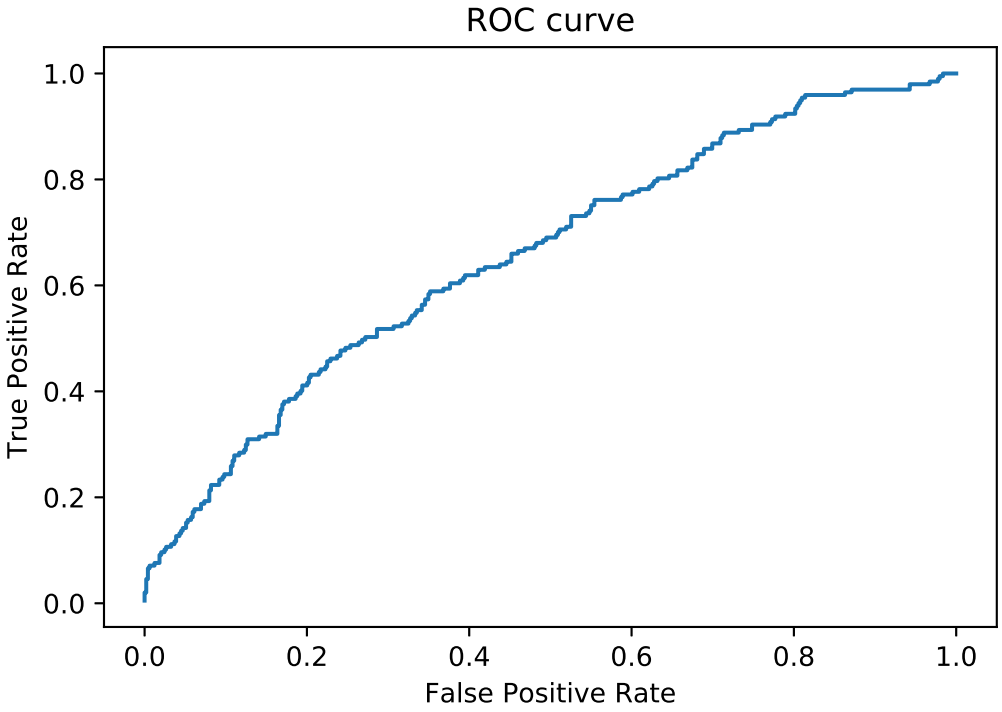

Figure 17: GBDT-6

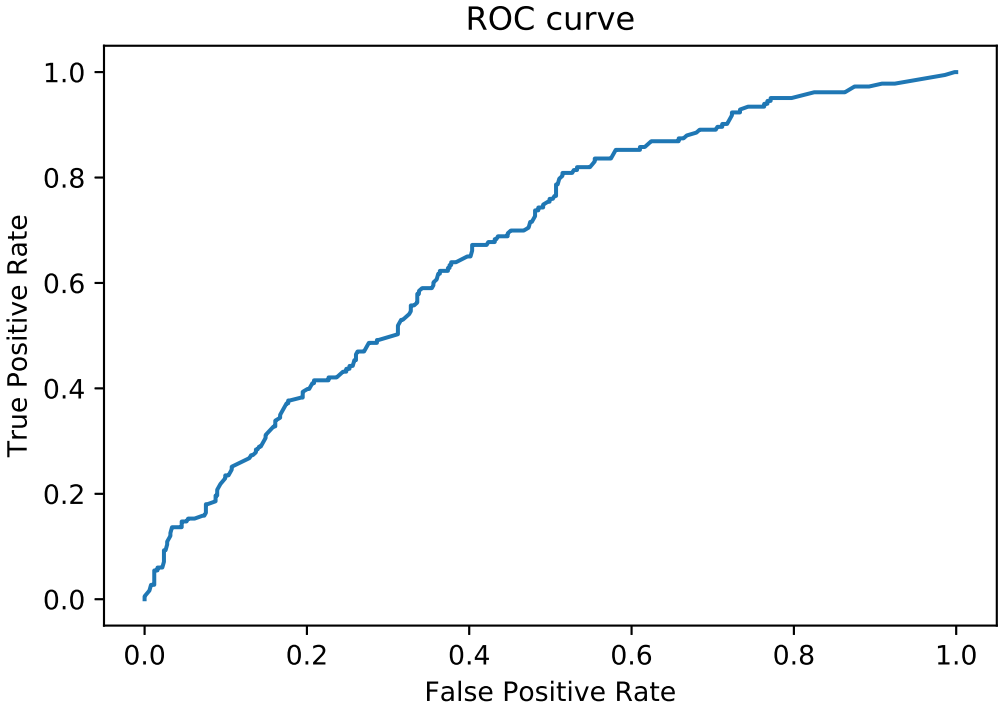

Figure 18: XGBT-1

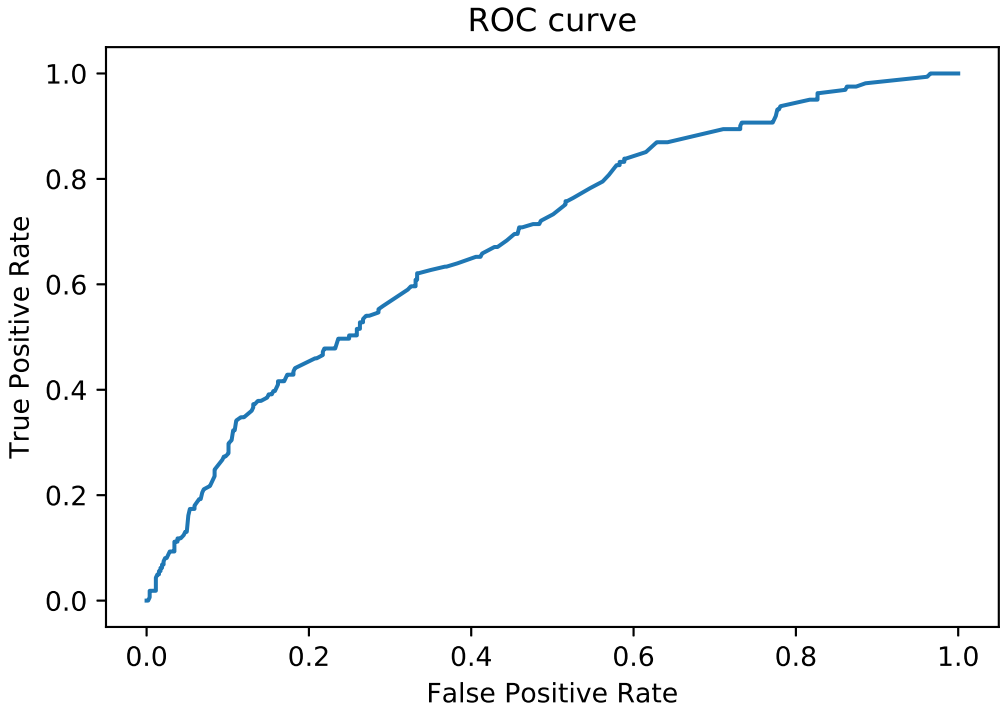

Figure 19: XGBT-2

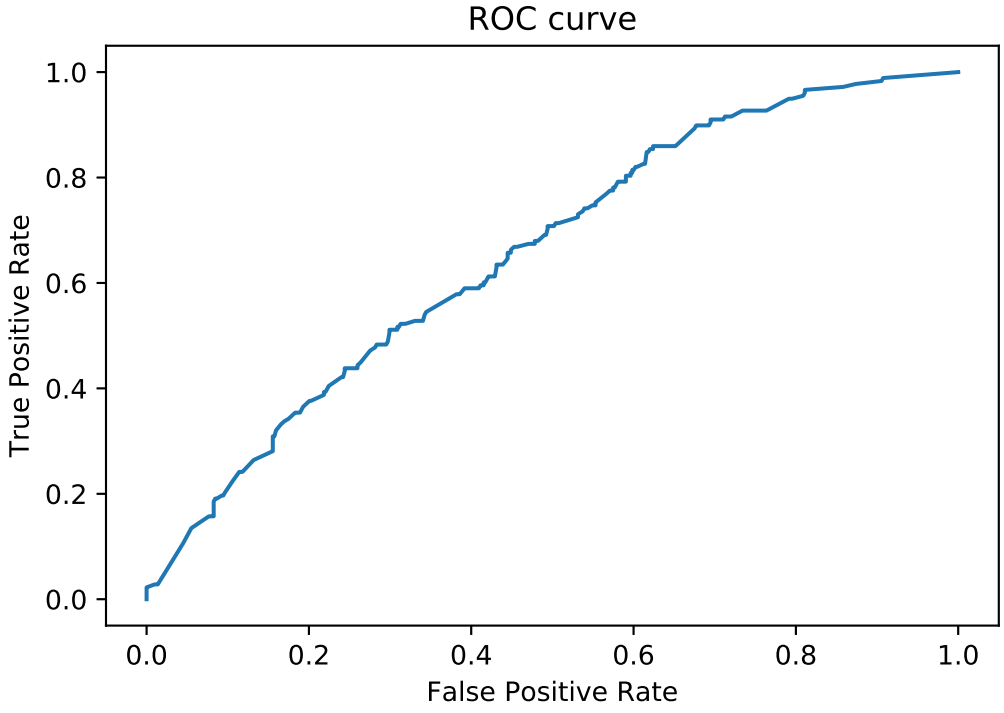

Figure 20: XGBT-3

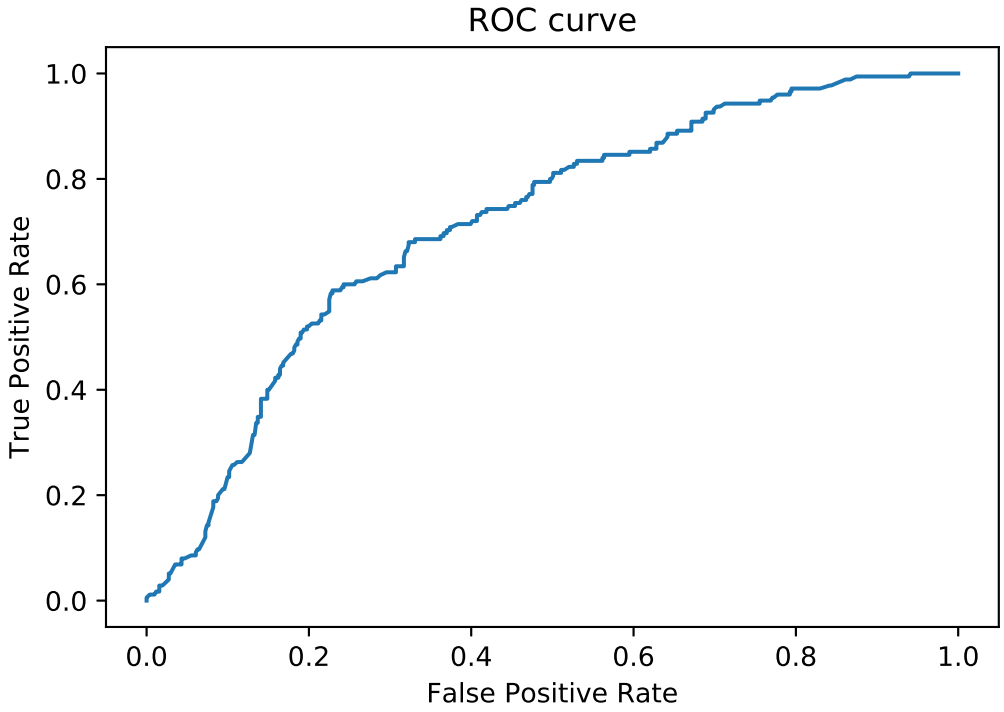

Figure 21: XGBT-4

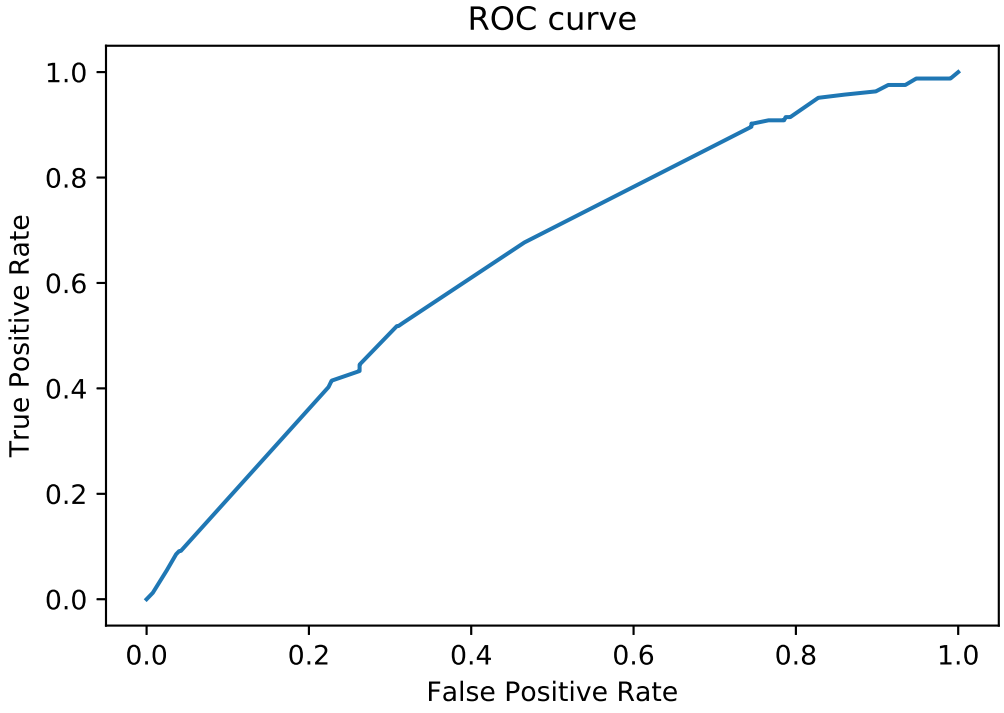

Figure 22: XGBT-5

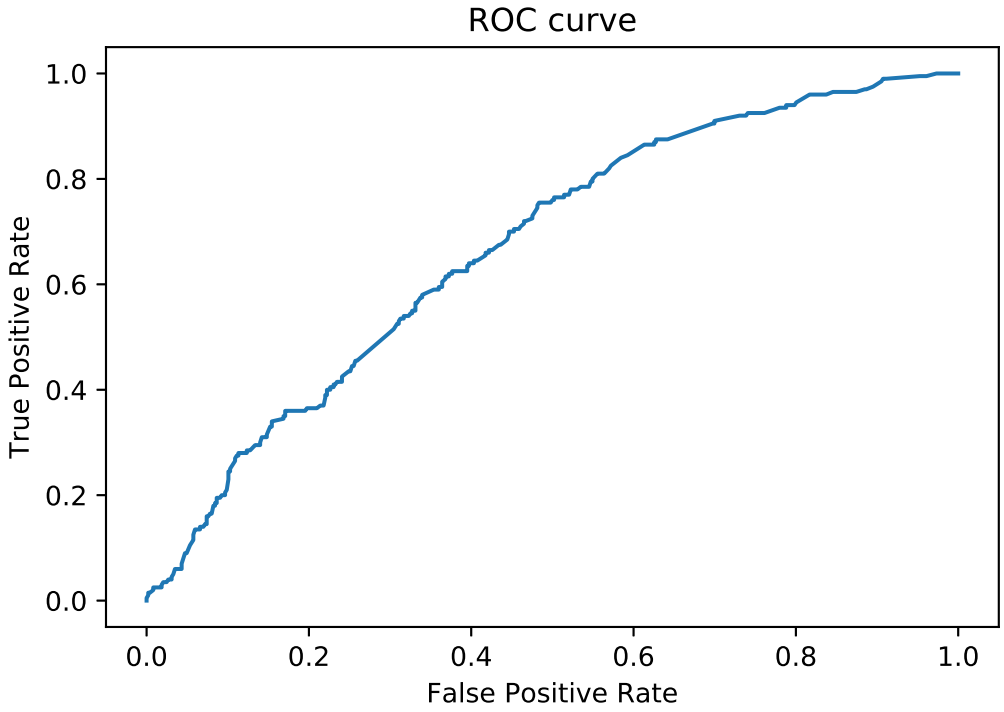

Figure 23: XGBT-6

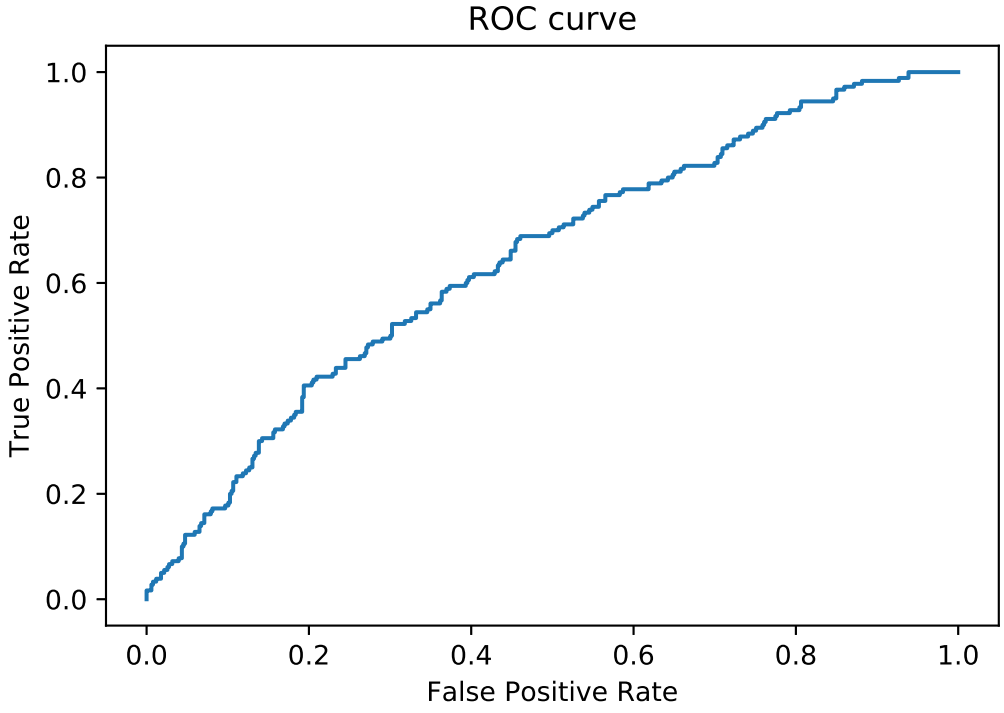

Figure 24: l_2-logistic-1

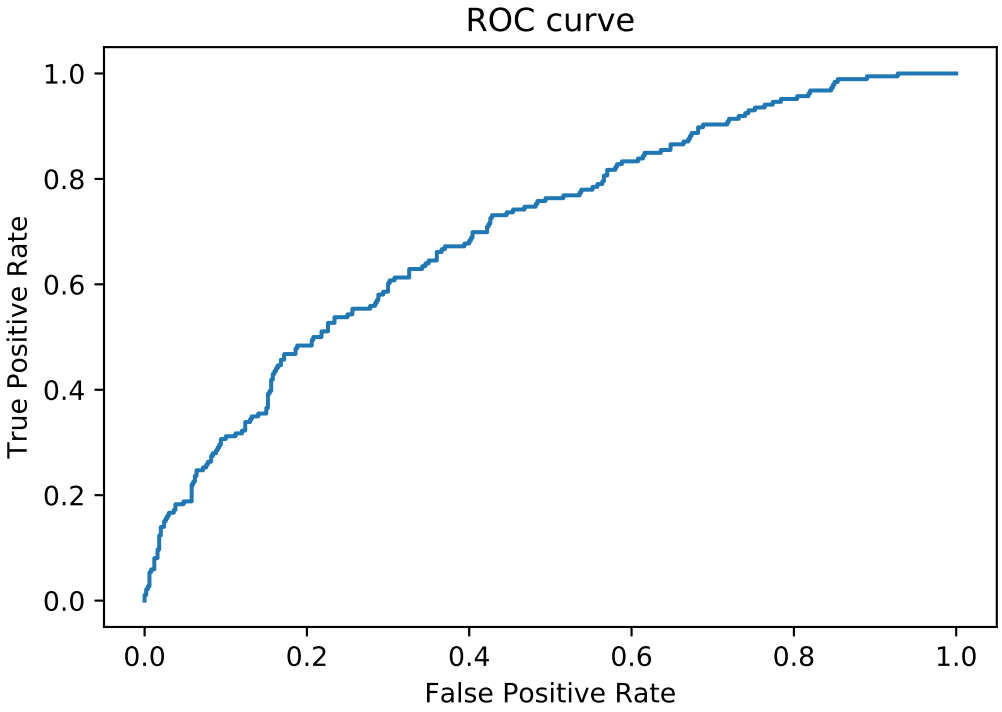

Figure 25: l_2-logistic-2

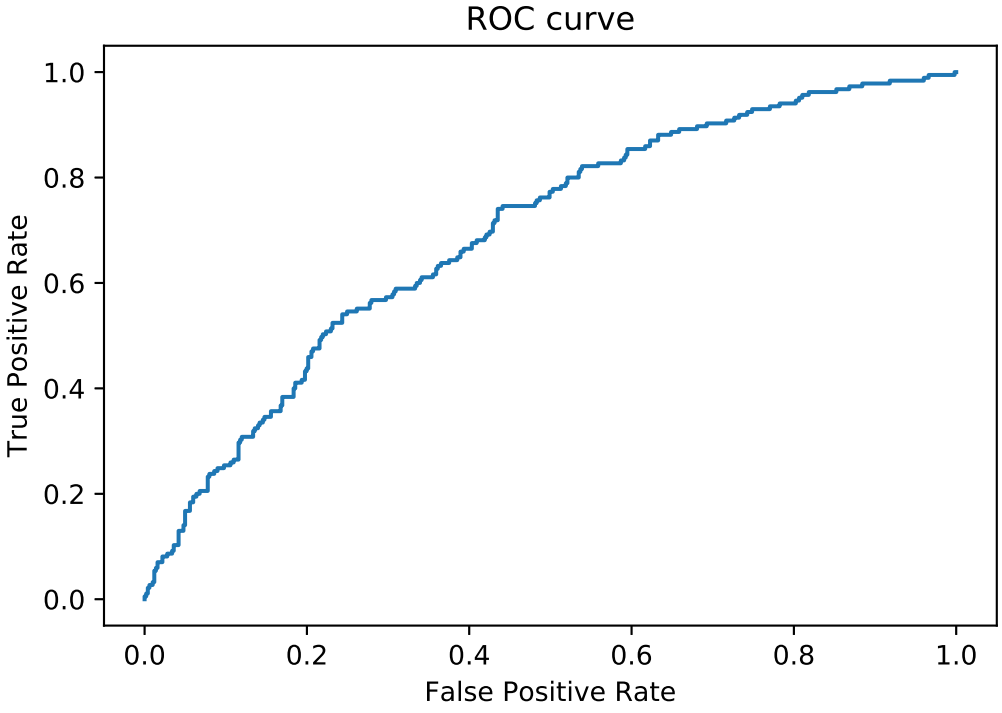

Figure 26: l_2-logistic-3

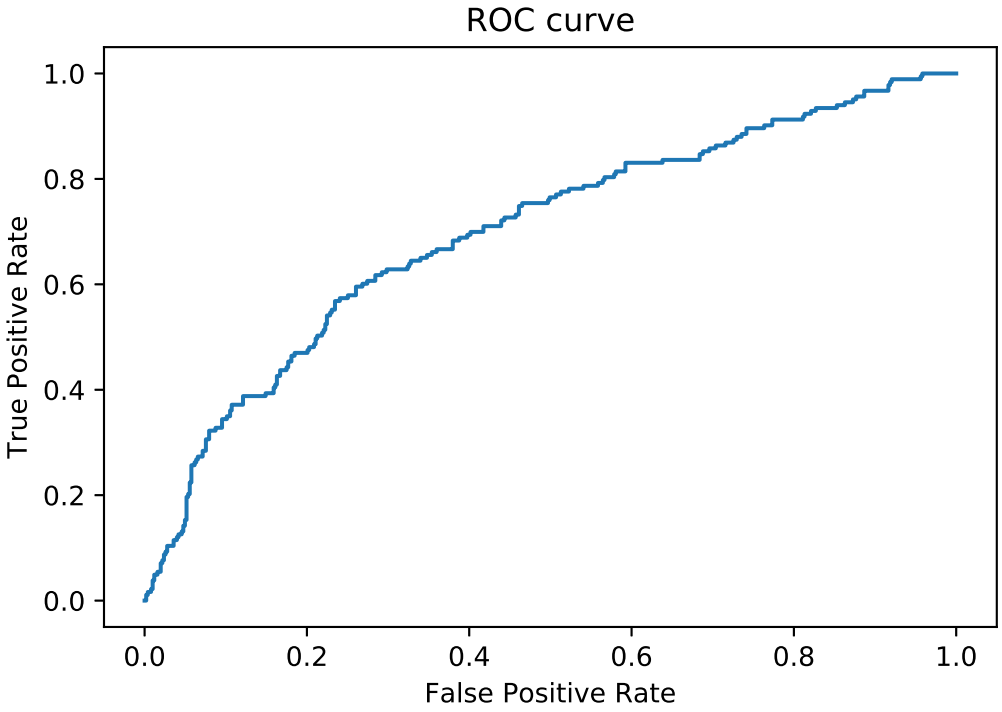

Figure 27: l_2-logistic-4

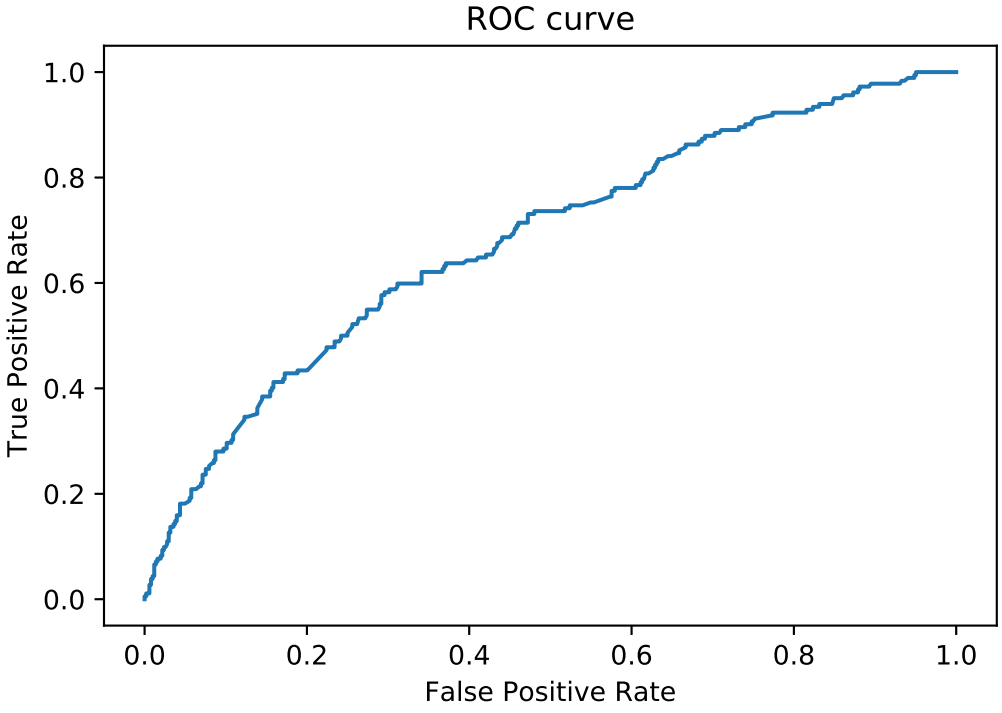

Figure 28: l_2-logistic-5

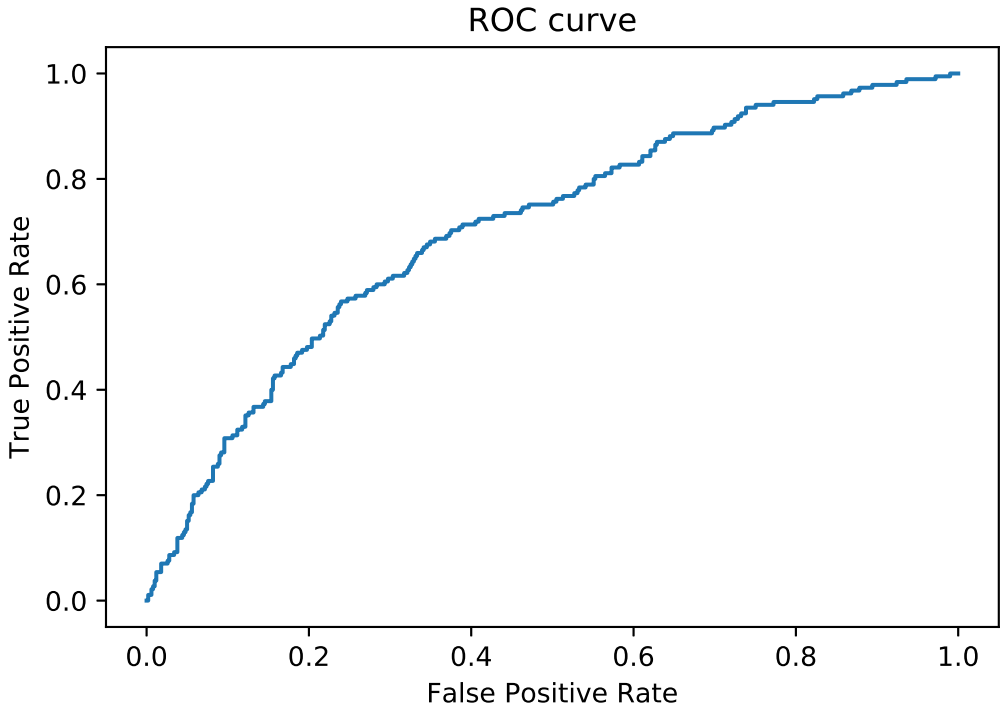

Figure 29: l_2-logistic-6

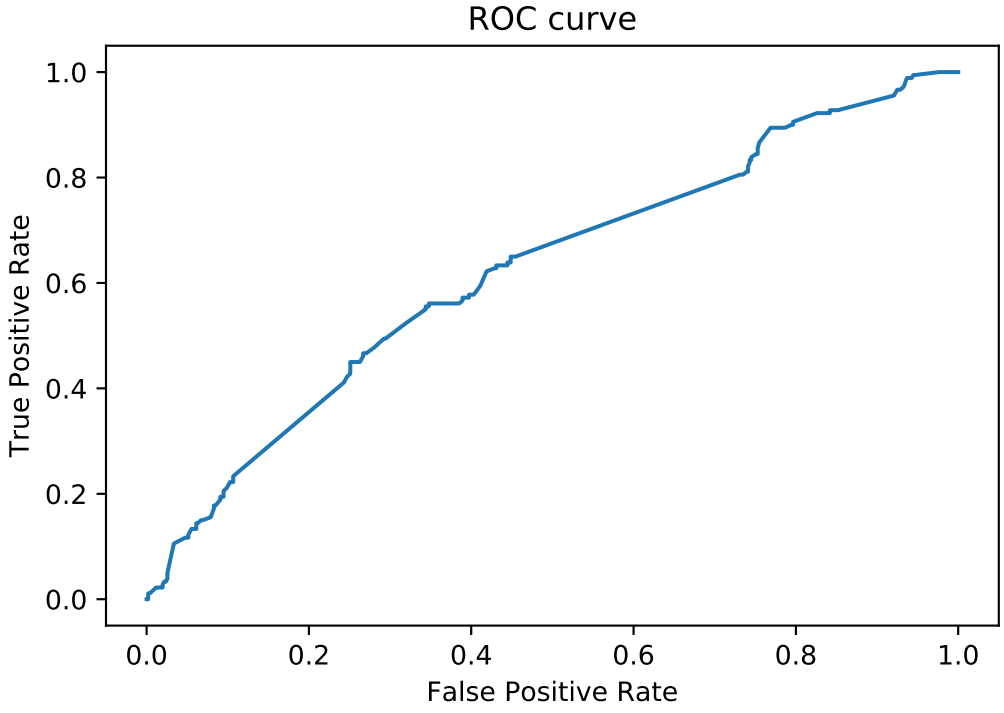

Figure 30: ELAS-1

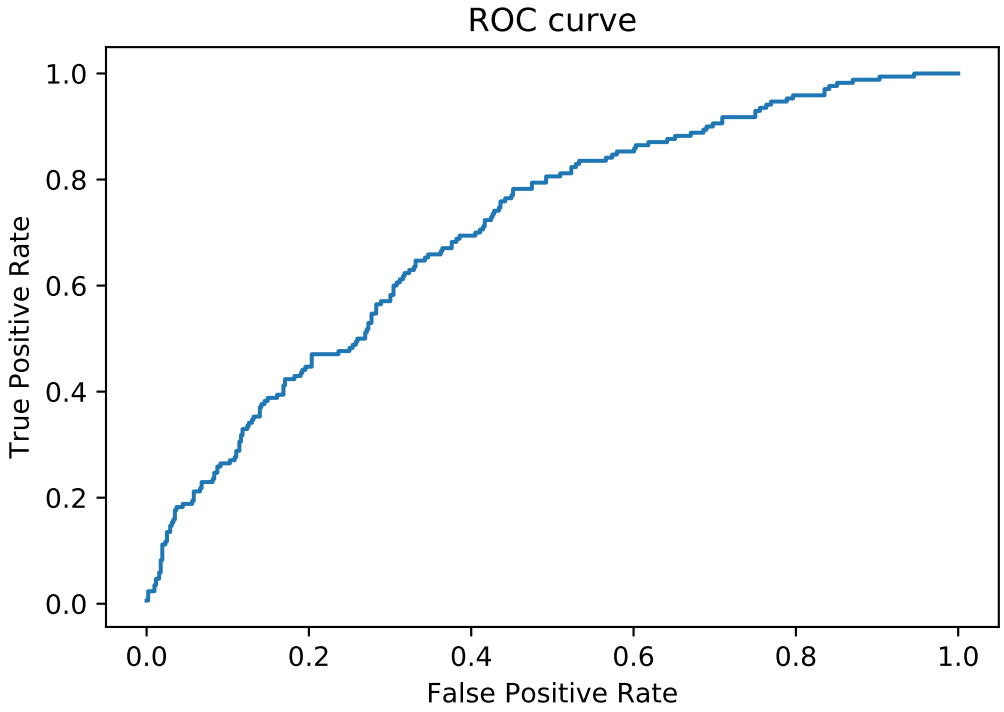

Figure 31: ELAS-2

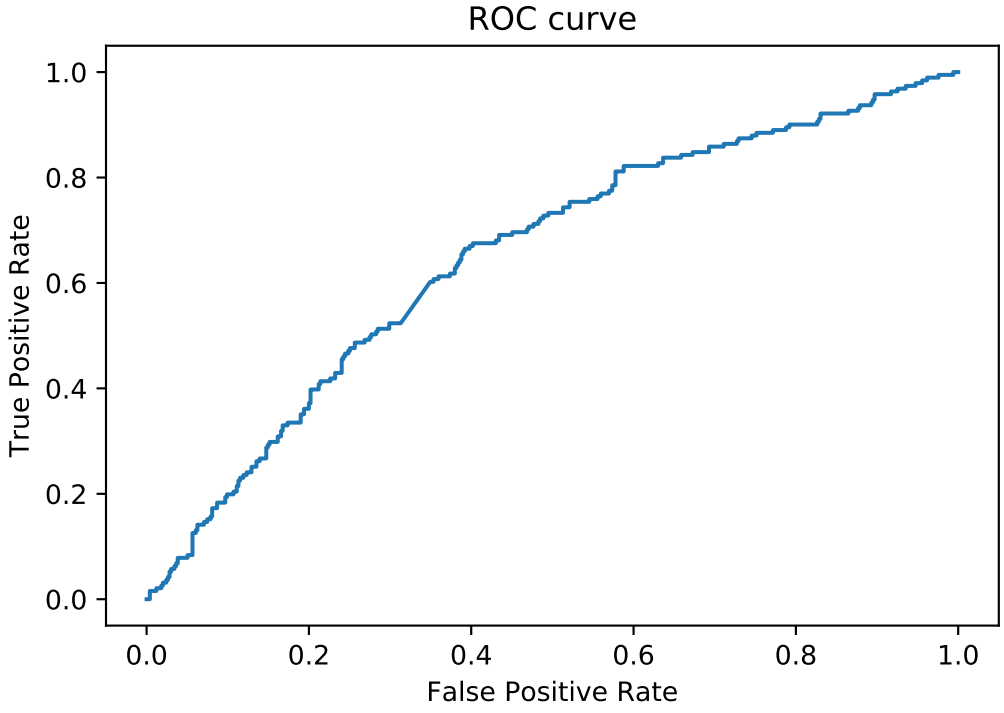

Figure 32: NN-1

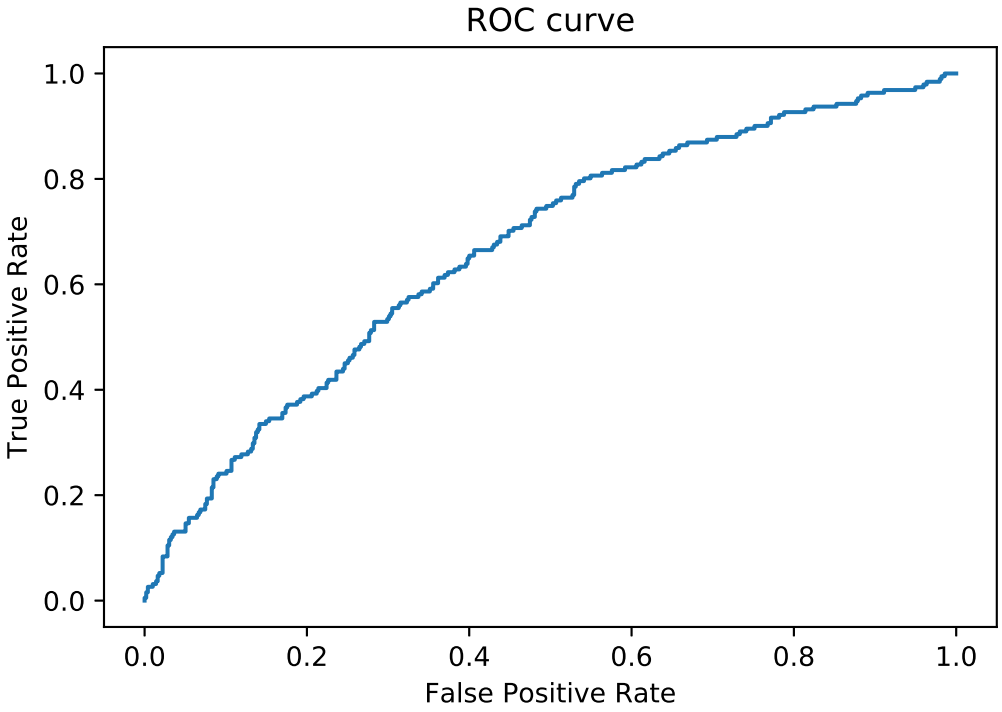

Figure 33: NN-2

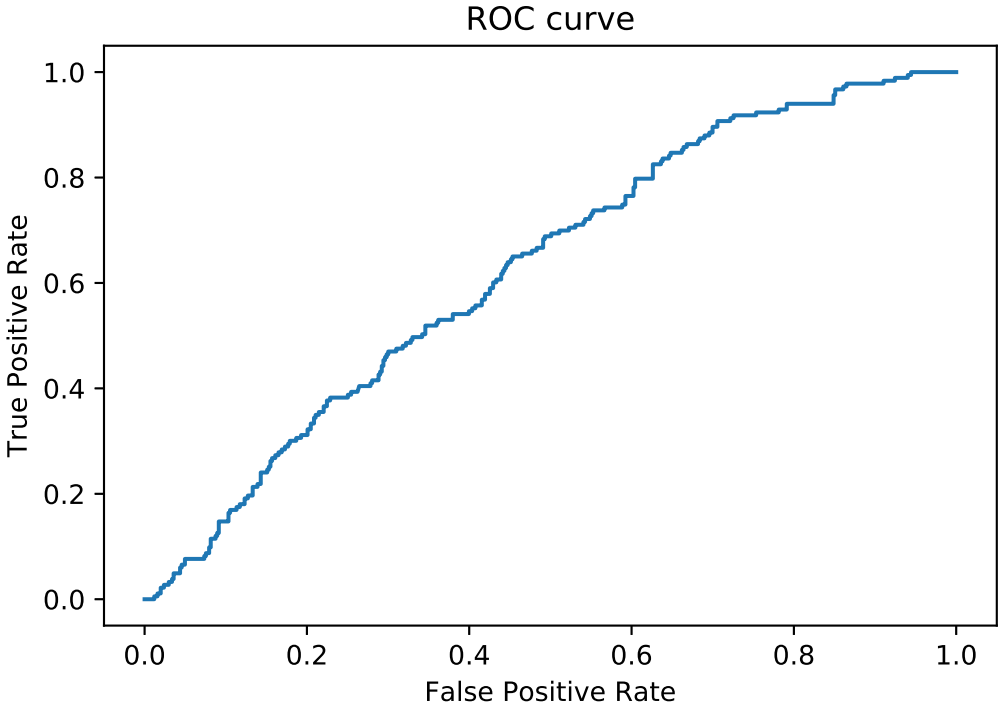

Figure 34: NN-3

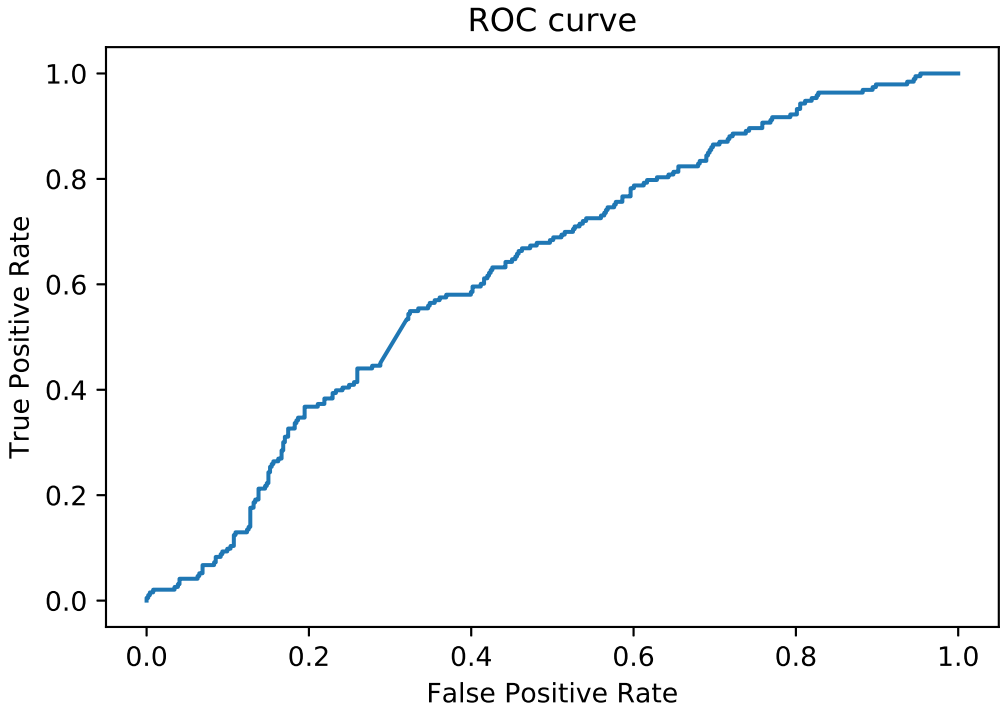

Figure 35: NN-4

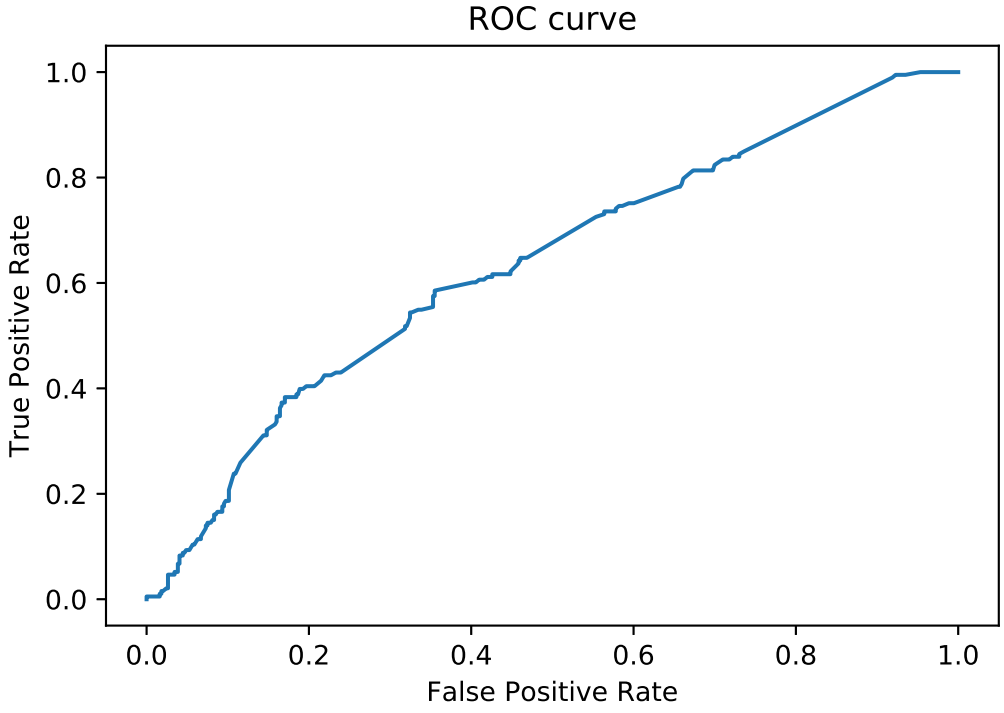

Figure 36: NN-5

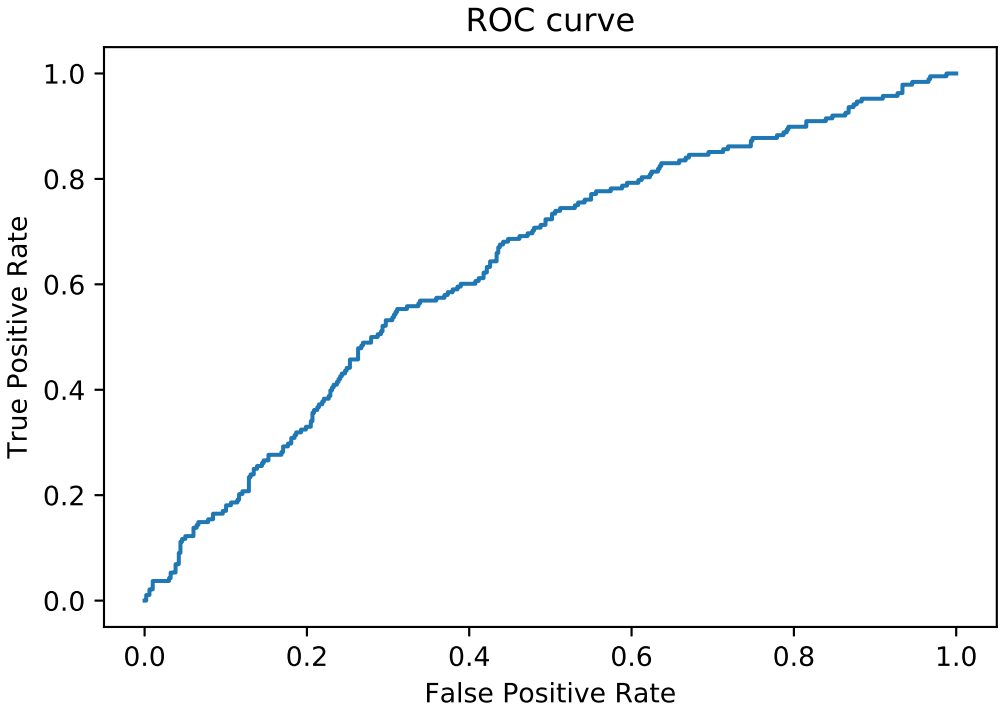

Figure 37: NN-6

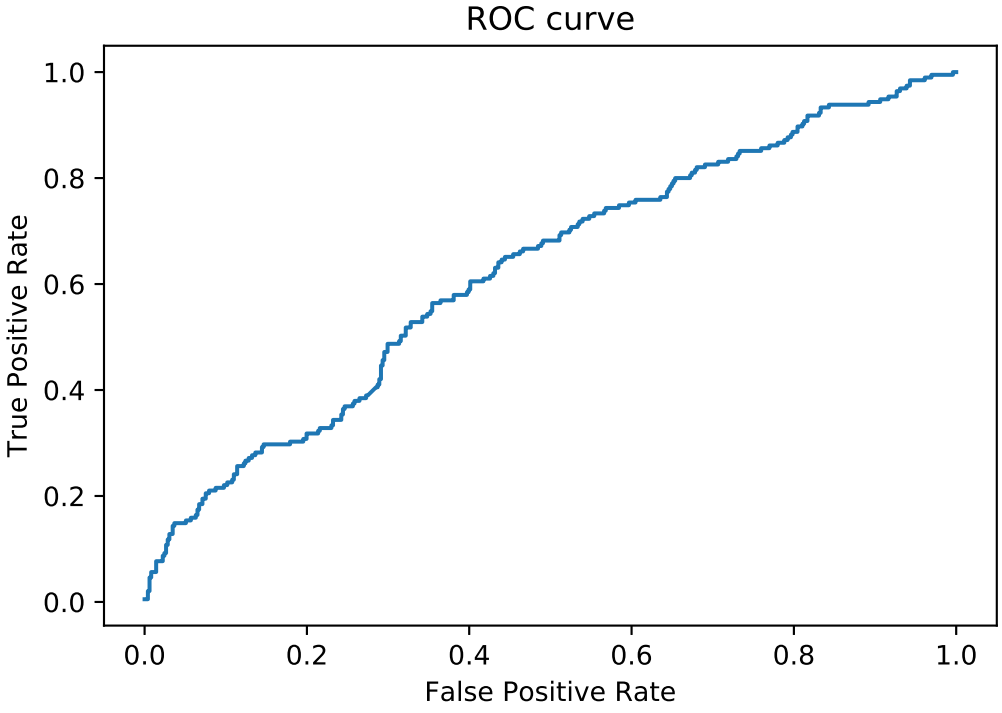

Figure 38: SVM-1

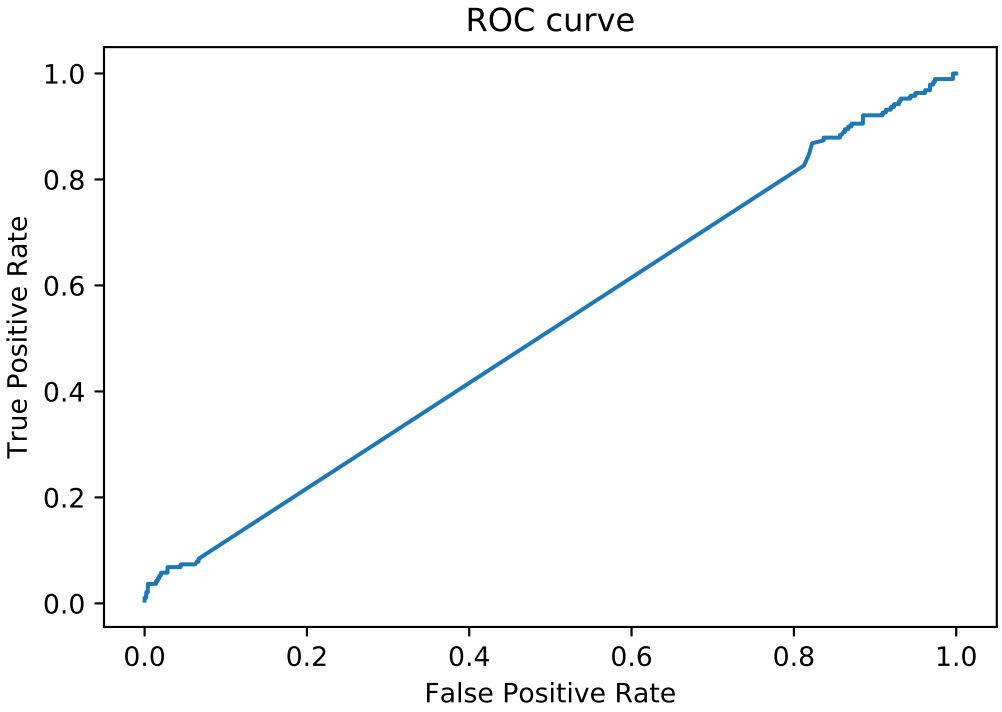

Figure 39: SVM-2

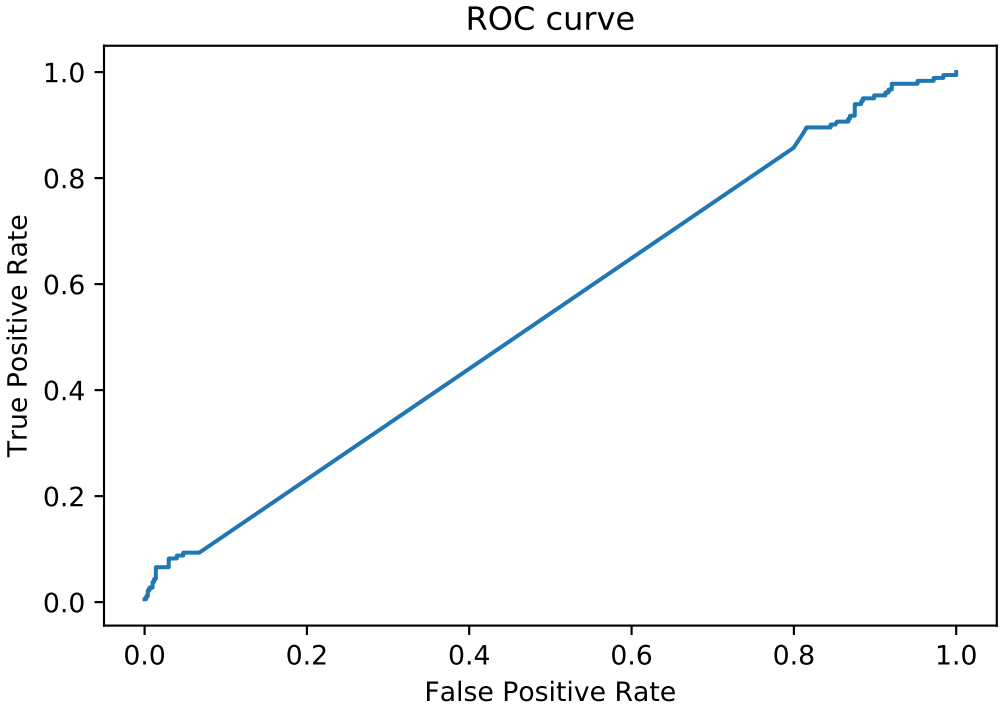

Figure 40: SVM-3

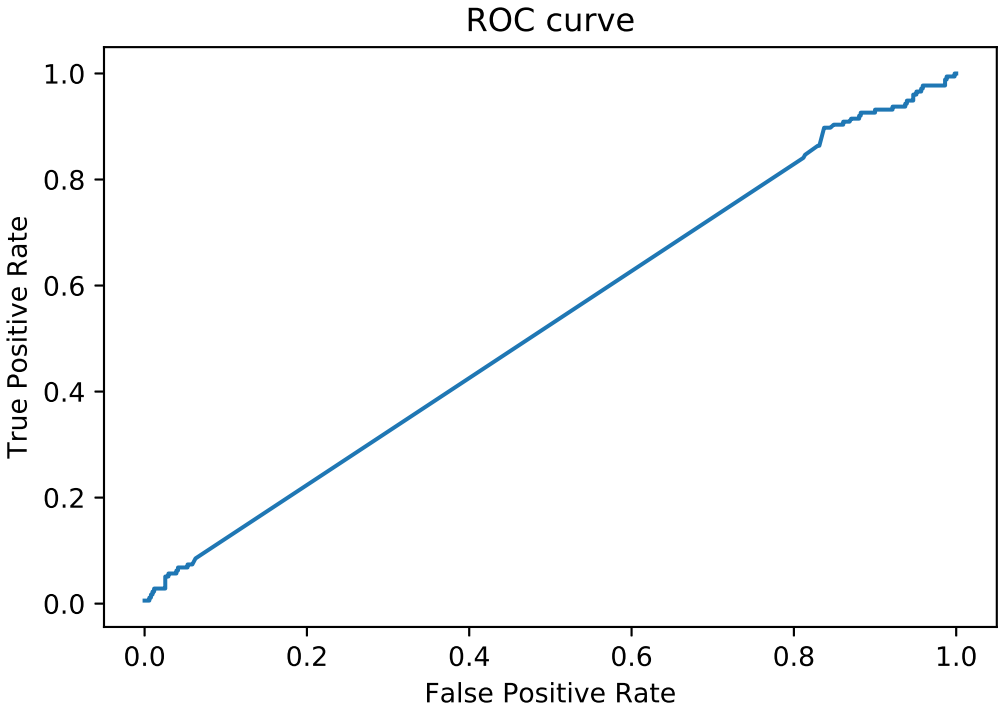

Figure 41: SVM-4

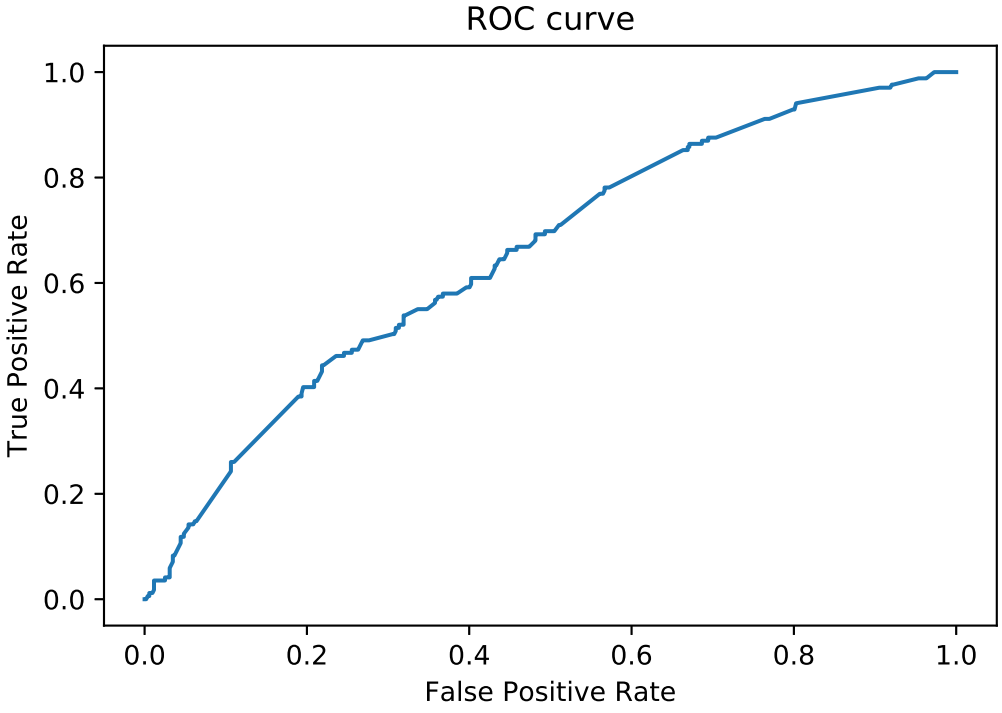

Figure 42: SVM-5

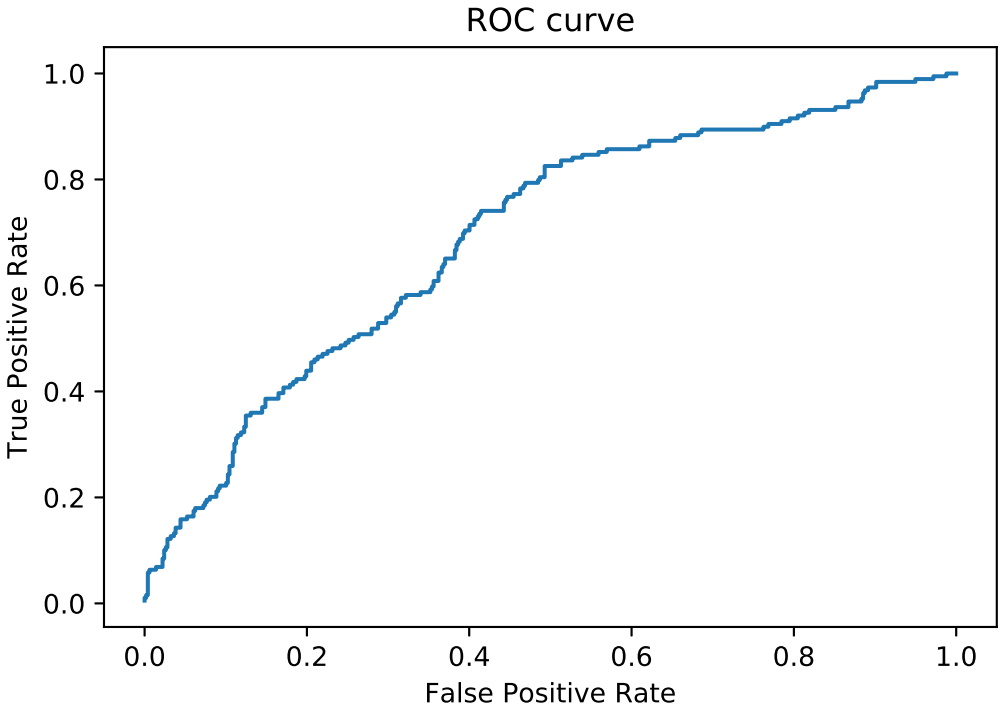

Figure 43: SVM-6

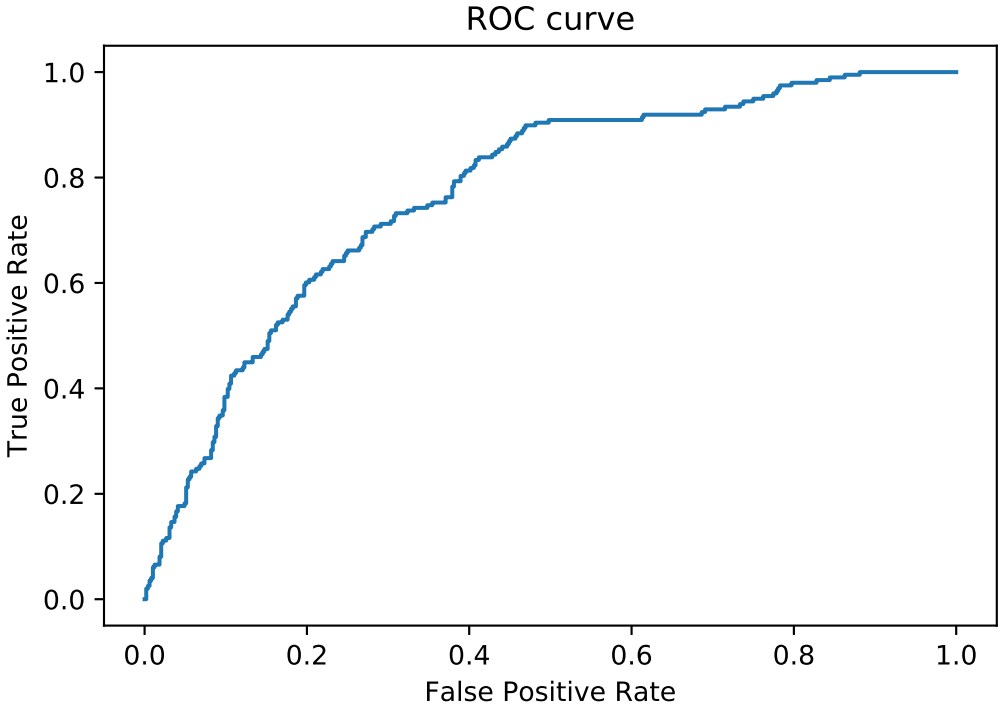

Figure 44: ultra-1

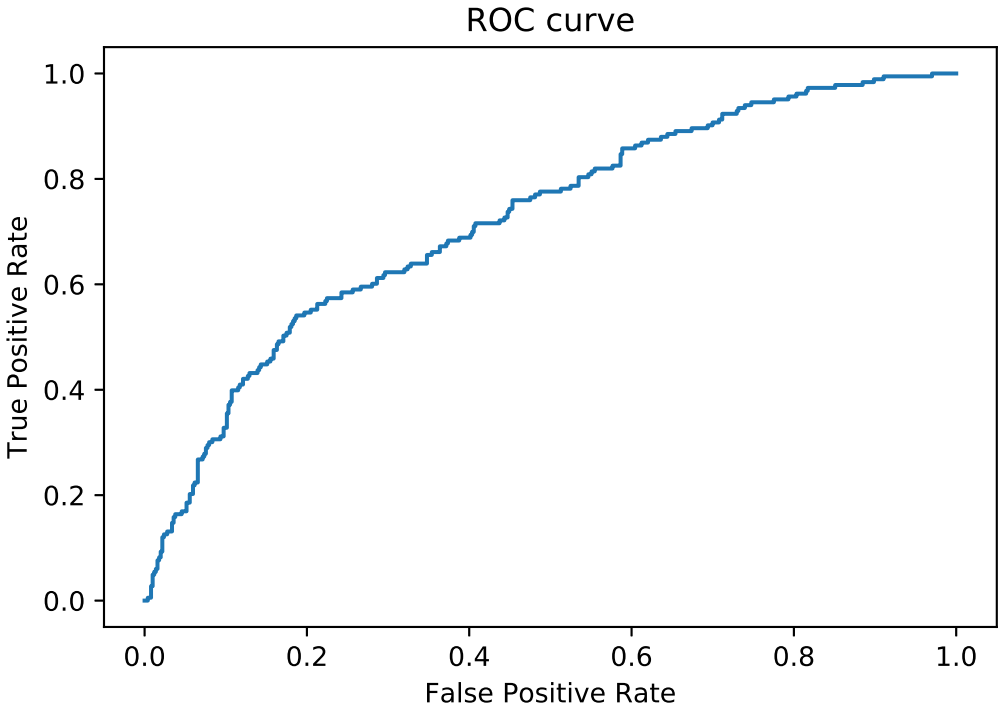

Figure 45: ultra-2

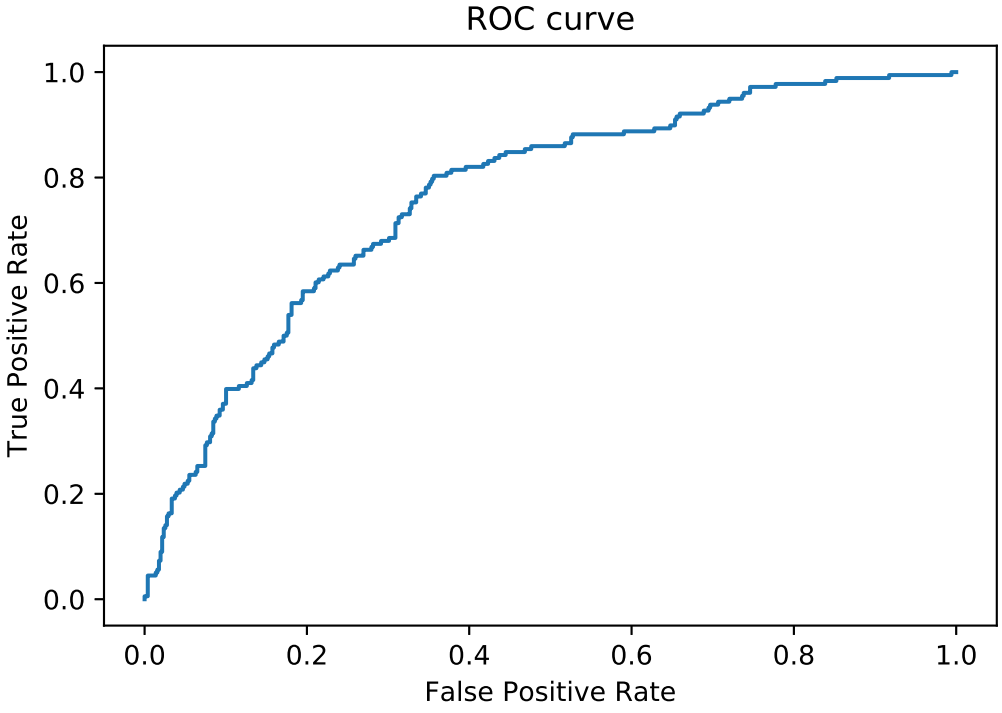

Figure 46: ultra-3

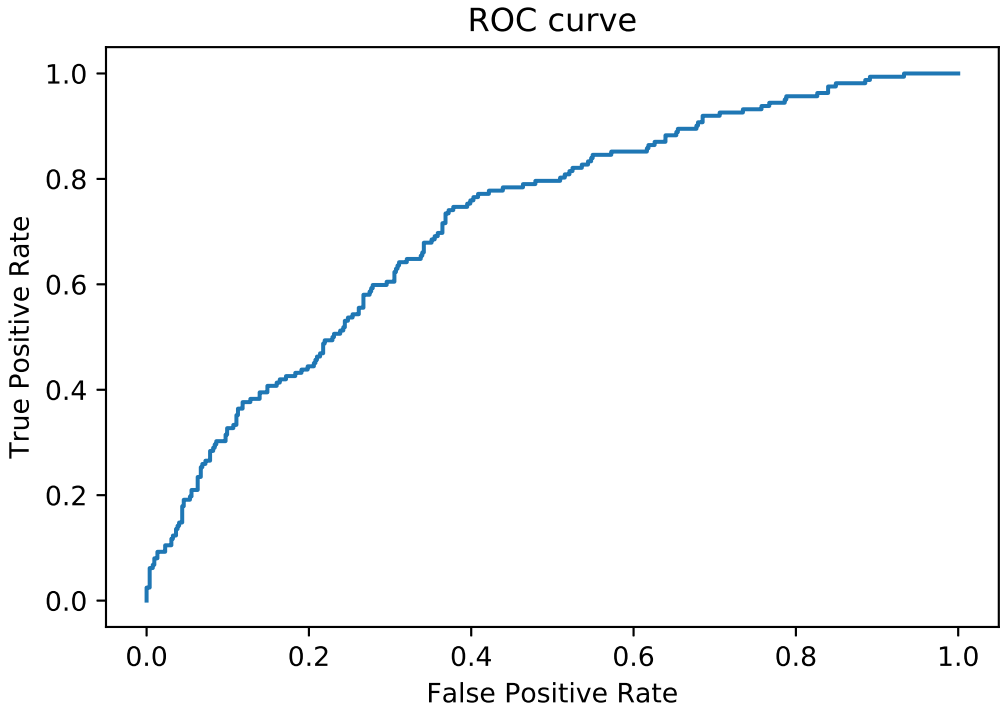

Figure 47: ultra-4

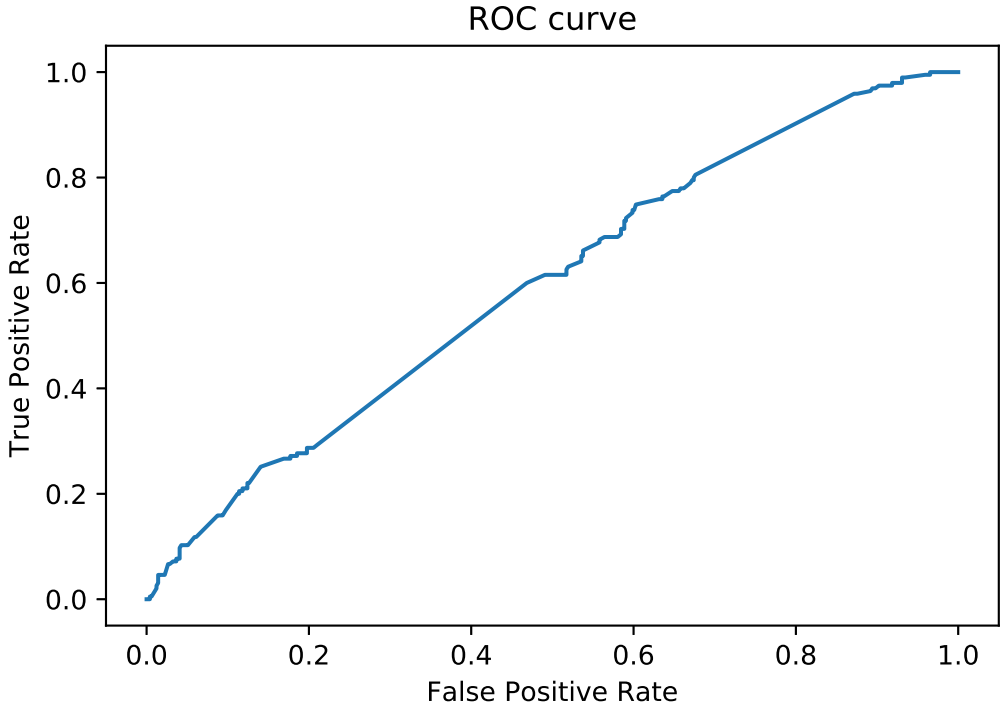

Figure 48: ultra-5

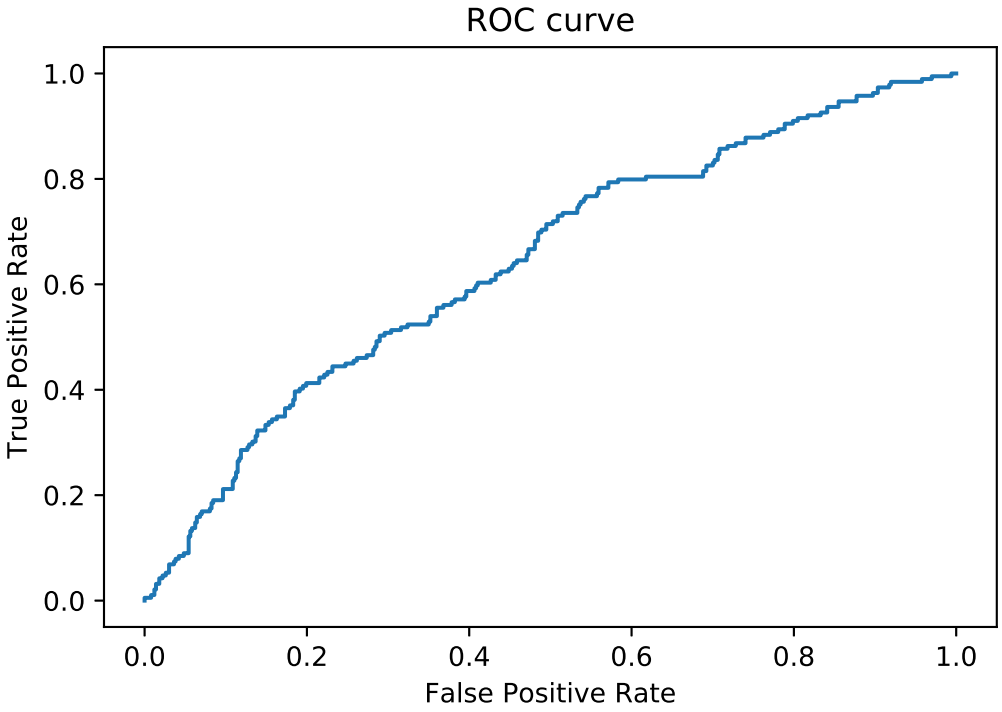

Figure 49: ultra-6

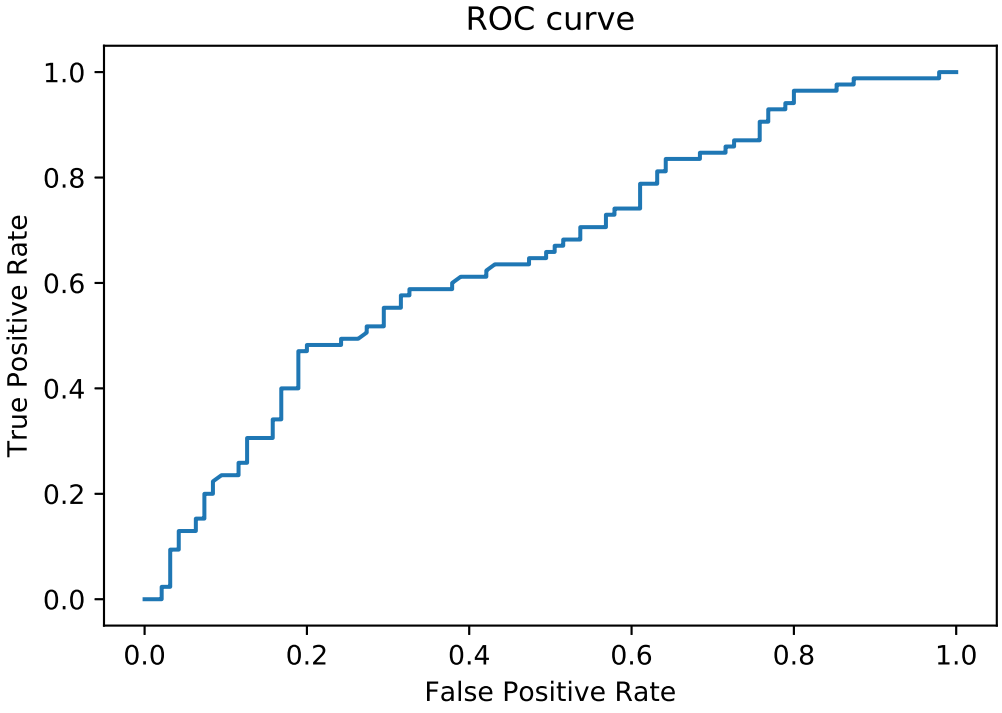

Figure 50: RF-Can

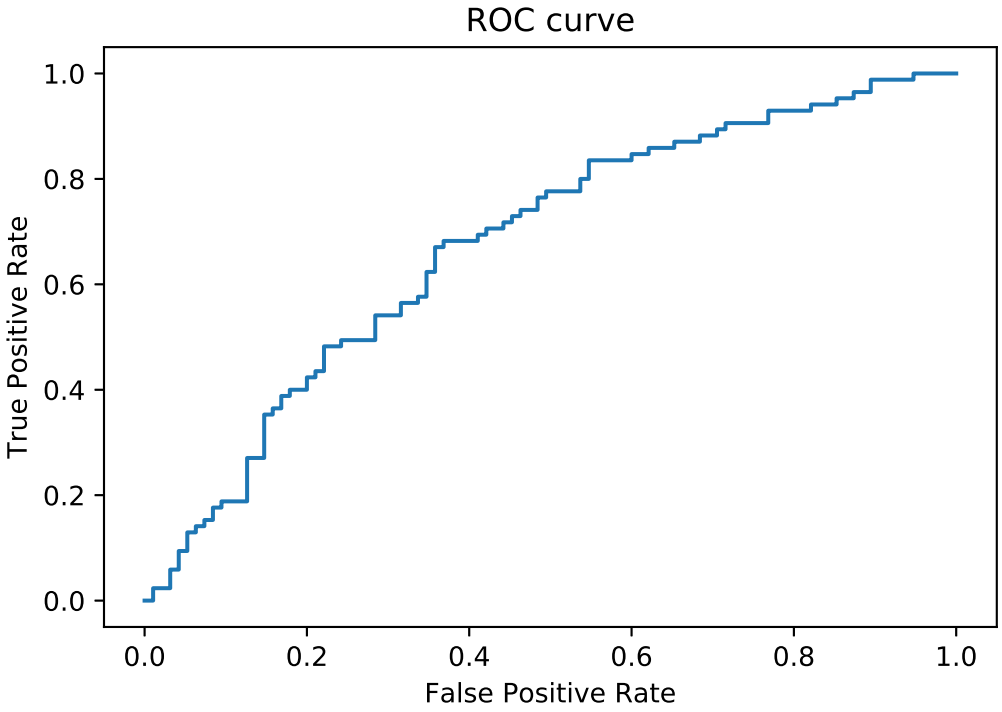

Figure 51: GBDT-Can

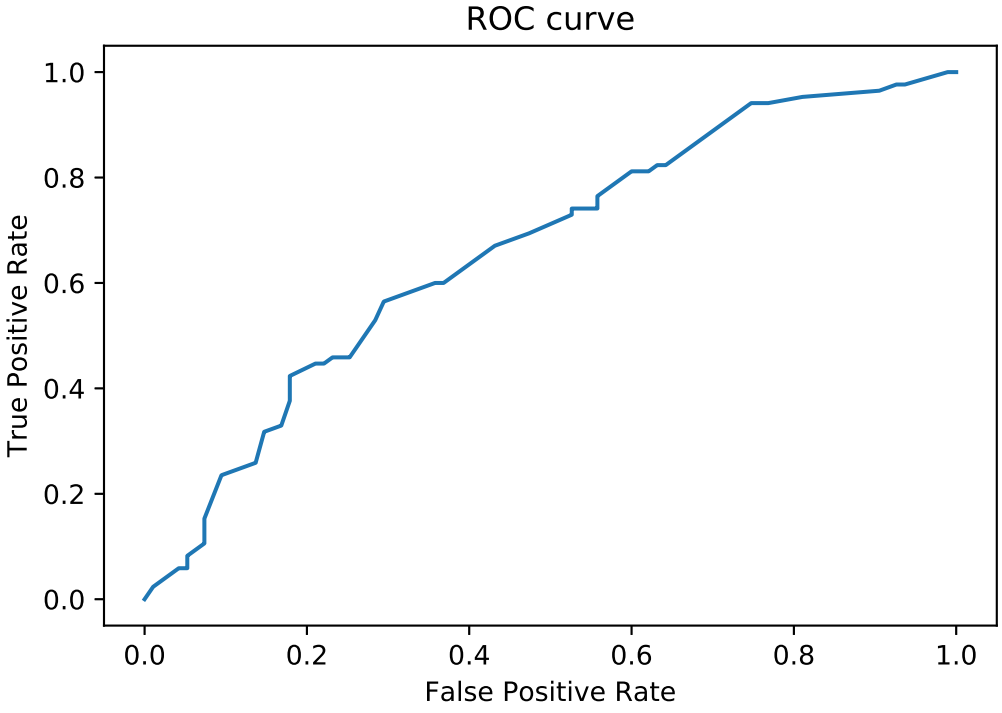

Figure 52: XGBT-Can

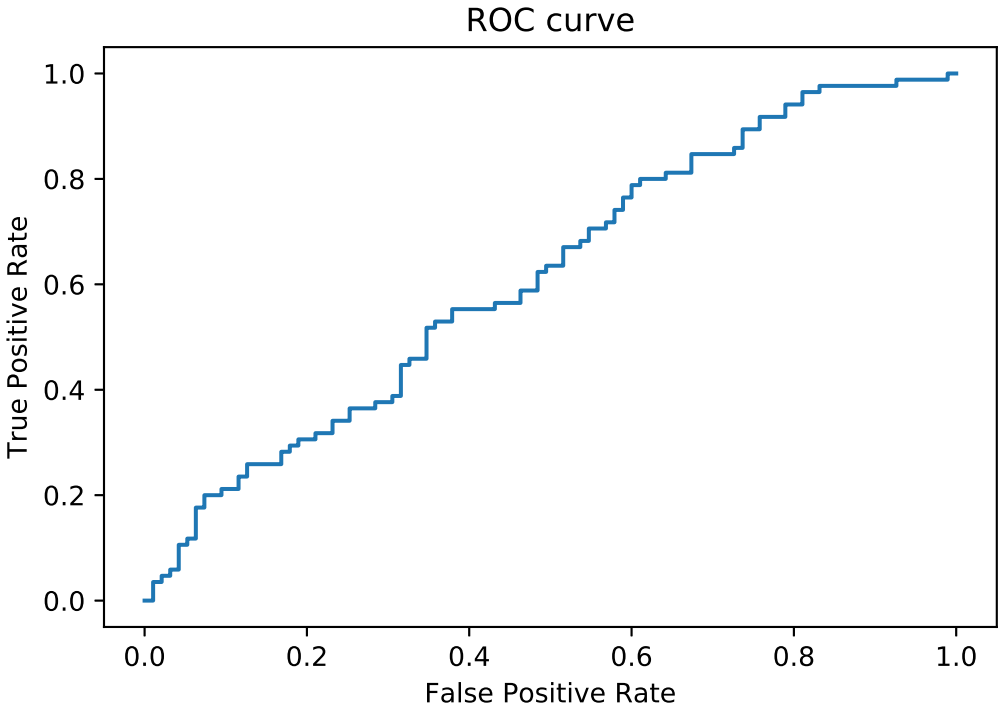

Figure 53: l_2-logistic-Can

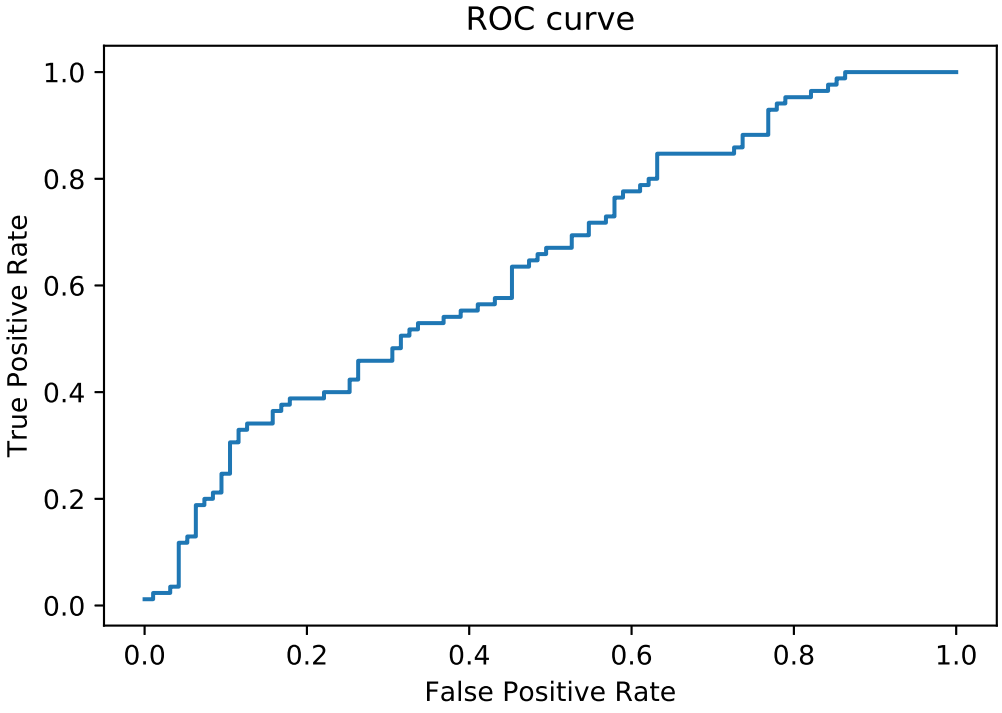

Figure 54: ELAS-Can

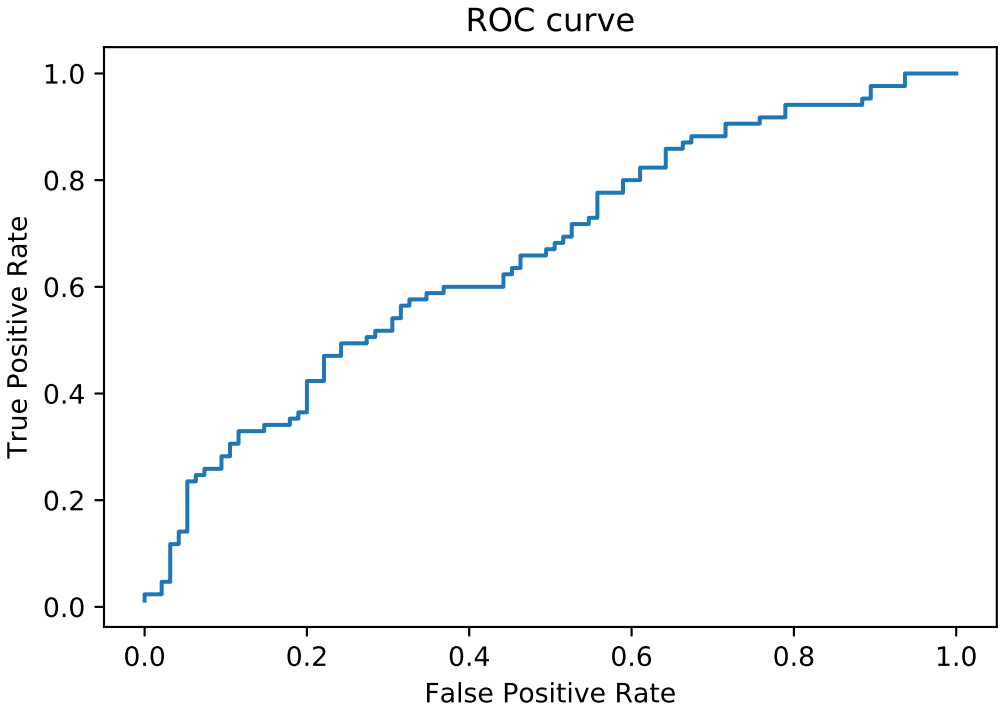

Figure 55: NN-Can

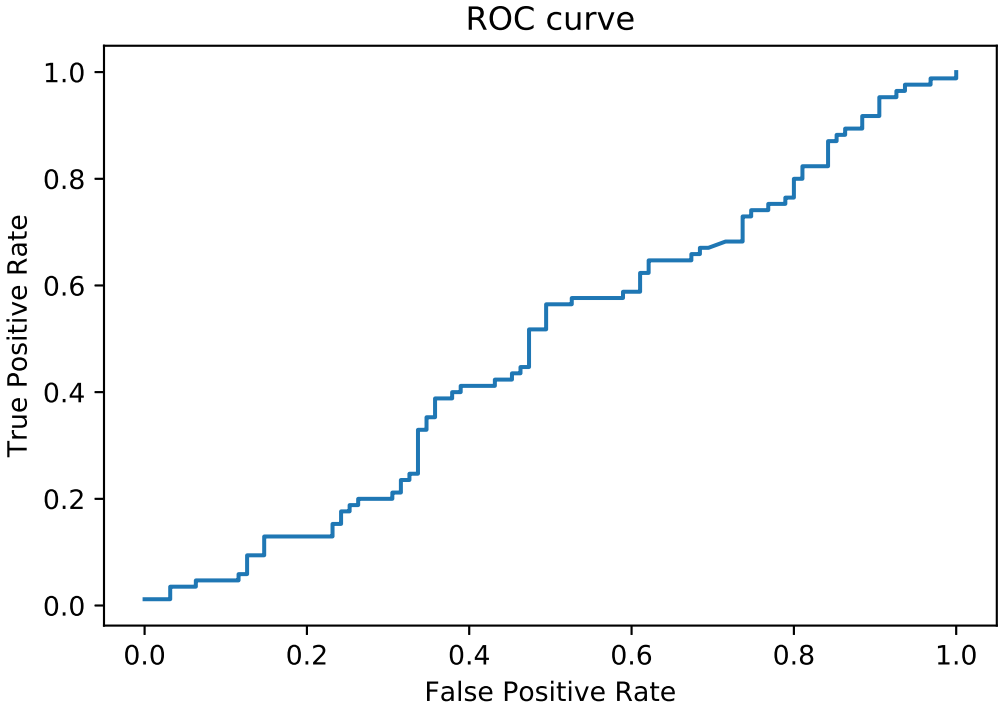

Figure 56: SVM-Can

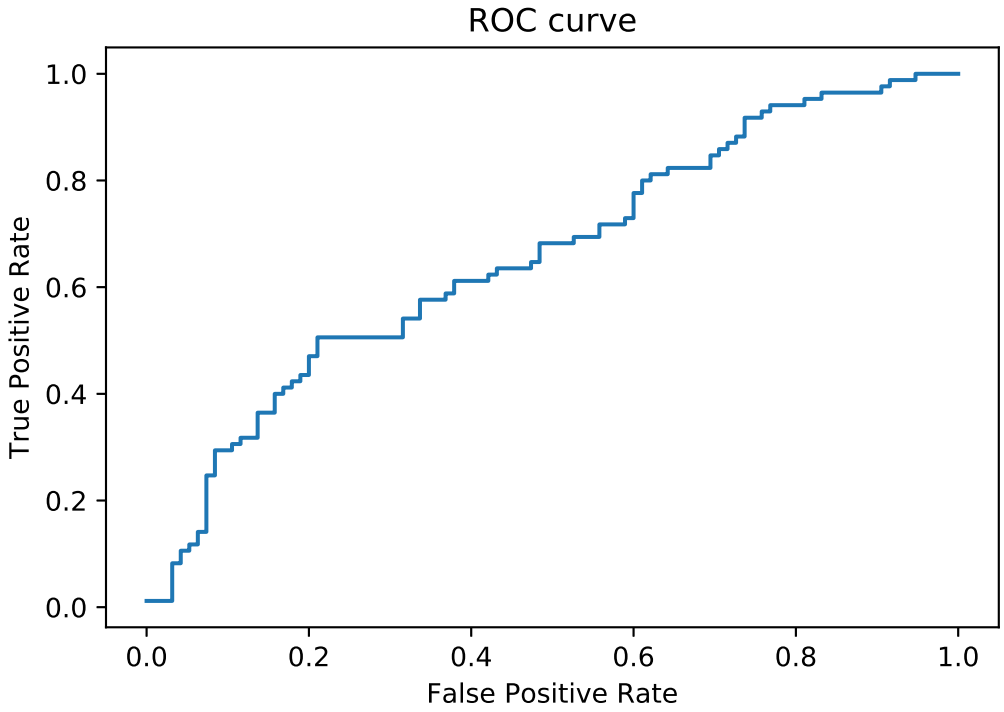

Figure 57: ultra-Can