To Be, or Not to Be, an Exoplanet

DS-GA 1001 - Fall ‘20 Term Project

Authors:

Anthony Chen - ac8480@nyu.edu

Sharder Islam - sti208@nyu.edu

David May - dmm9812@nyu.edu

Daniel Tang - dt1483@nyu.edu

Business Understanding

Astronomy has long been an important human endeavor. Regardless of how valuable it was previously thought to be, humans have had a constant and insatiable curiosity for the cosmos as early as our earliest recorded history. In the modern day, space exploration is largely funded and conducted by government agencies, with a few exceptions. This is likely due to the nature of progress in space exploration being a public good as opposed to a private good. Despite this, problems in space exploration are very much business ones. In private companies, the primary business goal is to maximize profits; whereas, in government agencies, the primary business goal is to maximize funding. In order to maximize profits, private companies are incentivized to create products that provide value to consumers. For government agencies to maximize funding, they are incentivized to create products that provide value to the *government* and by extension, to the people. One of the top priorities of government agencies such as NASA is the search for Earth-like planets in the quest for extraterrestrial life. This priority likely aligns well with the interests of many researchers at NASA, but more importantly, this priority lines up well with the interests of the general populace. NASA obtains its funding from the government, which obtains its funding from the people in the form of tax dollars. In order for NASA to maximize its funding, it must draw a collective interest of the people, and provide sufficient evidence of progress.

One of the most immediately apparent steps towards achieving the above goals is the search for exoplanets, which are planets outside of our solar system. In order to find Earth-like planets with potential life, it is necessary to first find the planets themselves. In NASA's quest for exoplanets, it searches for, and identifies numerous celestial objects as candidates. Following identification, NASA measures various attributes, such as the size and surface temperature, for each of these candidates. Each of these potential planets then undergo an intricate, multi-step process involving multiple scientists who perform a number of various verifying calculations to determine if the object is indeed an exoplanet or if it is just a false positive. (Landau, 2020)

In this research study, we propose a classification model that takes as its input the various attribute measurements of a celestial object, and outputs a predicted label of whether or not the celestial object is an exoplanet. Since the current state of the art process described previously requires a great deal of manual work from many different scientists, this model could potentially reduce costs significantly in the classification of celestial objects, if well-fledged out, making room for funding in other projects. Additionally, automating the classification process also expedites the process overall, which can potentially lead to faster progress in the search for Earth-like planets. This can consequently engender greater approval from the people and from the government to expand funding.

As the current state of the art process is done manually, it is inevitable that automating the process with machine learning may forego some accuracy. However, each step of the manual process is primarily reliant on the data, so we believe that the accuracy forgone may not be as significant as first appears (Landau, 2020). Of course, this potential dip in accuracy must still be addressed when the model is deployed, and will be discussed in the Deployment section. Next, we will discuss in detail the data at hand, and how we used it to create the prediction model, as well as examine how accurate our model is.

Data Understanding

The dataset used for this project is the NASA Kepler Objects of Interest table, obtained from a Kaggle dataset. There is no leaderboard scoring for this Kaggle dataset. It contains data collected by the Kepler Space Observatory satellite. This data includes 9564 entries of objects of interest, which are classified as confirmed exoplanets, candidate exoplanets, or false positives. The false positives make up 53% of the entries, while the confirmed constitute 24%, and the remaining 23% are candidates. We use the status of being a confirmed exoplanet as the classifier, as determined by NASA literature. Each entry has statistics gathered by the satellite, such as name, planetary radius, mass of the star, comments, and other scientific observations. There are 49 of these features. However, not all of these features are easily interpretable, nor relevant for modeling.

The 14 features we chose for modeling were those that we believe most accurately described the objects of interest through machine interpretable numbers or flags. The descriptions of each are as follows:

period: the orbital period in days of the object around its star. *impact*: the sky-projected distance between the center of the stellar disc and the center of the planet disc at conjunction, normalized by the stellar radius. *duration*: The time during which any part of the planet obscures the disc of the star. *depth*: The fraction of stellar flux lost at the minimum of the planetary transit. *prad*: The radius of the planet. *teq*: Approximate temperature of the planet in Kelvin. *insol*: Equilibrium temperature, which depends on the stellar parameters such as temperature and radius. *model_snr*: the time for the object to pass directly between the star and the observer (normalized by the mean uncertainty in the flux during the transits). *steff*: Photospheric temperature of object's sun. *slogg*: log of acceleration of gravity at surface of star. *srad*: Radius of object's star. *ra*: Right Ascension, or spatial angle along the equator of the planet's sun. *dec*: Declination, or spatial angle above the equator of the planet's sun. *kepmag*: kepler-band, size of energy emission detected.

The 29 columns we chose to ignore and why we chose to ignore them are as follows:

fpflag_nt, *fpflag_ss, fpflag_co, fpflag_ec:* Flags for Not Transit-Like, Stellar Eclipse, Centroid Offset, and Ephemeris Match Indicates Contamination. We ignored these because they are summary variables, composed of other features. This avoids collinearity and redundancy. *koi_tce_delivname:* TCE delivery name corresponding to the TCE data federated to the KOI. This feature does not include any comprehensible data. *koi_tce_plnt_num:* Redundant numbering of data. *koi_time0bk:* Time corresponding to the center of the first detected transit in Barycentric Julian Day (BJD) minus a constant offset of 2,454,833.0 days. This feature introduces unnecessary time-dependance to the data, because an object of interest should be independent of the time it was discovered. *22 Error columns:* These columns would increase model complexity for a small gain in performance.

A source of selection bias in this data is that larger objects are easier to find with a telescope. As telescopes get better in the future, there may be smaller objects that are not accounted for by this model.

Data Preparation

After the process of exploratory data analysis as well as data understanding of our columns, we know that Kepler objects of interest (KOI) are identified when they display at least one transit-like sequence in the Kepler time-series photometry. This is tied with the planetary transit hypothesis, which is the dimming of the star that occurs when a planet orbits in front of it. ‘Pdisposition’ is the most probable disposition that the Kepler data analysis has towards the exoplanet candidate, where it is a ‘CANDIDATE’ or ‘FALSE POSITIVE’. Following this nominative process, in the NASA archives, a specific ‘Disposition’ is a categorical indicator for an exoplanet candidate, which are ‘CANDIDATE’, ‘CONFIRMED’, and ‘FALSE POSITIVE’. The ‘CONFIRMED’ exoplanets are candidates that have already been approved by NASA scientists from the candidate batch. According to the archives, “False positives can occur when: 1) the KOI is in reality an eclipsing binary star, 2) the Kepler light curve is contaminated by a background eclipsing binary, 3) stellar variability is confused for coherent planetary transits, or 4) instrumental artifacts are confused for coherent planetary transits" (NASA, 2013).

In the given data, a KOI disposition is determined by the KOI score that takes on a value between 0 and 1 that indicates the confidence in that particular categorical label. NASA calculated this value based on a Monte Carlo technique so that it is tantamount to the fraction of iterations where the Robovetter, NASA’s own robotic decision-making software, yields a disposition of ‘CANDIDATE’ (Coughlin, 2017). For candidates, a higher value indicates more confidence in its disposition, while for false positives, a higher value indicates less confidence in that disposition.

Initially we decided to engineer a feature, “new scores,” which was between -1 and 1 to indicate the confidence of being a confirmed exoplanet, but decided to continue with a feature engineered binary target variable. This was because the ‘new scores’ would be continuous, and it would not make our target result a firm, identifiable classification of being a valid Kepler object. To examine the target variable distributions, we grouped the dispositions along with their respective KOI scores, and found the means of ‘CANDIDATE’ to be 0.90, ‘CONFIRMED’ to be 0.96, and ‘FALSE POSITIVE’ was roughly 0.014. Thus, this gave us direction to use the ‘Disposition’ column to choose our binary target variables, and designate a value of 1 if it was a ‘CONFIRMED’, and 0 if it was a ‘FALSE POSITIVE’. We wanted to avoid using the ‘CANDIDATE’ labeled exoplanets as our targets because it could potentially become a false positive. As for the ratios of the dispositions, 53% of the entries were ‘FALSE POSITIVE’, 24% were ‘CONFIRMED’, and 23% were ‘CANDIDATES’.

After looking at all the possible data fields, and excluding some columns, only 14 were used as our actual features. As previously mentioned, we removed the ‘fpflag_’ variables because they heavily influenced our prediction results in our initial test run of feature importance. A ‘not transit-like’, ‘stellar eclipse’, ‘centroid offset’, or ‘ephemerics match indicates contamination’ flags were highly correlated with the KOI score, which we want to avoid during training. This would make it difficult to determine the individual impact of collinear predictors on response. In addition, ‘Koi_tce_delivname’, ‘koi_tce_plnt_num’, and ‘koi_time0bk’ were all dropped because they were extraneous labels that had no underlying connection with our target variables. Finally, all error columns were removed from our dataframe as well.

To further prepare the 14 most important features generated from a decision tree model’s gini importance metric for the binary target variables we also cleaned the feature labels so that the ‘koi_’ was deleted. This allowed our column selections to be more clarified and less prone to typing errors. Finally, we selected all the relevant columns that we would train on starting from the ‘period’ index, and imputed missing values in each column with their respective column means. With our data preparation completed we could move onto modeling.

Modeling & Evaluation

The task at hand was to classify Kepler objects of interest based on the properties of the object and of its stellar environment. A plethora of features were available for use, but many were composite features, or features that encoded uncertainty in the given estimate. Extraneous variables were also dropped as previously specified. Ultimately, the goal was to predict which of these Kepler objects was a likely exoplanet and warranted closer examination. Using a binary target variable, multiple classification models were considered. Each model predicted whether a given Kepler object was either a "FALSE POSITIVE" or a "CONFIRMED exoplanet. Kepler objects in the "CONFIRMED and "FALSE POSITIVE" statuses were previously in the status "CANDIDATE". Further examination through other tests allowed the investigators to distinguish whether these objects were truly exoplanets or just false positives. Kepler objects that remained in the status "CANDIDATE" were dropped. A series of classification algorithms were considered and the AUC from each model was compared to find the one that best predicts true exoplanetary status. Considering the importance of true positives, the AUC metric is well suited to evaluate models that make predictions for this classification.

Considering the nature of the task, the first algorithm chosen was the decision tree model, a classification algorithm suited for such prediction. The decision tree was trained on 75% of the dataset and tested on the remaining 25% after completing the data preparation steps previously outlined. This was performed with our binary target variable using the out-of-the-box hyperparameters for the algorithm. The model was used to predict the classes for the remaining 25% of the data, i.e. the test data. The AUC score for this model was calculated and stored for later comparison with the other models. In addition, feature importances were derived from the model to understand which of the features lent the most explanatory value to the model overall. The feature with the highest importance for the decision tree was the planetary radius. This aligns with our intuition of the physics behind celestial formation: smaller objects would not have enough mass, and consequently not have sufficient gravity, to be habitable.

A logistic regression was also performed on the same training data. The logistic regression is another classification algorithm suited to the task. The baseline model for this logistic regression was performed with out-of-the-box hyperparameters as well. Similarly, a support vector classifier was trained on the same data but with its kernel specified as “linear”.

All other hyperparameters were left on their out-of-the-box settings for this algorithm. After training all three classification algorithms, the models were used to predict the class for each Kepler object, and AUC values for each model were compared to determine the best predictor of exoplanetary status. At this stage, the best predictors were both the logistic regression and the support vector machine, since they had excellent AUC values. The AUC values for each of these base models can be found in Figure 4. To improve on the model, a grid search was performed on the three algorithms to find the best hyperparameters. For the logistic regression, the parameters that were searched were the value for C, i.e. the regularization strength, and the norm used in penalization. For the support vector classifier, the value of C and the kernel were modified. In the polynomial kernel, the degree was searched up to the third degree. The best model generated was the decision tree with the ideal minimum samples per leaf at 0.005, and the ideal value for the minimum samples split at 0.015. The AUC for this best estimator from the grid search was 0.95. The model risked being overfit due to such low values for the hyperparameter, and more candidates for modeling algorithms were considered.

To check for further improvements in predictive ability, ensemble classifiers were considered. This included the random forest, adaptive boosting, and gradient boosting classifiers. Each of these classifiers were trained on the same data as the previous three classifiers with all of the hyperparameters left at their default setting. In total, there were six classifiers considered but there were a total of 80 models considered as part of the earlier grid search. As expected, the ensemble classifiers were all better than the logistic regression, with the best being the random forest. Considering the best base model was the decision tree, it is not unexpected that the random forest was the best of the ensemble classifiers. Because the AUC value for this particular classifier was sufficiently high, a decision was made to forego a grid search on the ensemble classifiers. The AUC values for these ensemble classifiers can be found in figure 3.

To further elucidate the key features of the final model, Shapley Additive Explanations (SHAP) were leveraged. SHAP is an algorithm used to explain complex models using a game theoretic approach (Lundberg & Lee, 2017). It confirms that the planetary radius is the most important feature for each Kepler object, as previously confirmed by the decision tree importances. As shown in Figure 1**,** a low planetary radius was extremely predictive of the class but having a high planetary radius was not as predictive. The next most important feature was the orbital period, followed by the equilibrium temperature. Similar inverse predictive patterns were seen for these two features. Low values for these two features were predictive but high values were not. The results shown in Figure 2 give the feature importances according to SHAP separated by the classifier. The proportion given to Class 0 represents SHAP importance for objects determined to not be an exoplanet and the proportion for Class 1 represents the importance for those determined to be an exoplanet. The information gained from this SHAP analysis could be used by researchers without access to this model in order to classify future objects of interest.

With this random forest classifier, one can predict whether a given Kepler object of interest is truly an exoplanet. Scientists at institutions such as NASA, SpaceX, or even Blue Origin, can leverage a model such as this to identify celestial objects for further inquiry. Space missions have exorbitant costs so an accurate classification model will save both financial resources as well as the time of the professionals who work at these organizations. With increasing private interest in space exploration, the value of models is expanding past just only governmental space agencies such as NASA. This model can assist in narrowing the scope of which objects are considered for missions and therefore, reduce the resources needed to perform any in-depth investigations. With this utility in mind, the model needs to be precise, with as few false positives as possible. However, considering that the results of this model will still need to be reviewed by an astronomer, the penalty for too many false positives is not as dire as it might be for other classification tasks. The AUC value allows for comparison of the false positive rate to the true positive rate to choose the best algorithm, and it also relates to the business problem by maximizing precision in predicting our target variable with a binary separation.

Deployment

Despite the promising accuracy of our model tests, it is necessary that the model not be fully deployed immediately. The deployment of this data mining process can be installed and monitored in batches. Due to the nature of the task, confirmation of the model's classification cannot be done until it is investigated by a member of the research team with subsequent testing. Incoming objects detected by the Kepler telescope can be classified by the model and predictions can be stored until a later time. These predictions can then be confirmed or denied, and then the model can be improved with this new data. The model can continue to be evaluated in terms of its true positive rate and false positive rate in the form of the AUC. This iterative process will fine tune the model based on the features already considered. However, improvements in telescope technology or new scientific breakthroughs may lead to a paradigm shift that would require re-assessing the model from scratch, but the new model would still be informed by the same findings from this initial classification task. This may also lead to consideration of new features, or make previously considered features more important.

One possible issue that the firm should be aware of is that, since our model was trained and evaluated only on those data points that have already been either labeled confirmed or false positive by NASA, it is possible that the model will not perform as well on more ambiguous points which are still labeled as ‘CANDIDATE’. The ambiguity of these points may introduce some selection bias to our model. Another possible issue that the firm should be aware of is the delay in observations for objects that are further in light years than those that are closer, as it takes longer for devices to take measurements for objects that are further away. Our model could be used as a basis for further evaluation through the use of ensemble methods to determine if an exoplanet is habitable. If an exoplanet is indeed hospitable, this may lead to future inequity in accelerated space flight based on socioeconomic status or even other factors.

Wrong KOI identifications of exoplanets would not cost many resources and time if scientists want to study a potential planet further. We want to consistently retrain cyclically as new data emerges. Due to the innocuous nature of the task, there is not much risk associated with this proposed plan. The readings from the Kepler telescope are already available to the researchers, and the model assists in connecting the factors already collected.

Works Cited

1. Coughlin, J. L. (2017, May 25). NASA Technical Reports Server (NTRS). Retrieved November 30, 2020, from https://ntrs.nasa.gov/citations/20170009549
2. Landau, E. (2020, September 24). How do you find – and confirm – a planet? 10 things about the search for exoplanets. Retrieved December 02, 2020, from https://exoplanets.nasa.gov/news/1524/how-do-you-find-and-confirm-a-planet-10-things-about-the-search-for-exoplanets/
3. Lundberg, S. M., &amp; Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Retrieved December 01, 2020, from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
4. NASA Exoplanet Science Institute. (2013). Data Columns in Kepler Objects of Interest Table. Retrieved November 30, 2020, from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html
5. NASA. (2017). Kepler Exoplanet Search Results. Retrieved December 01, 2020 from https://www.kaggle.com/nasa/kepler-exoplanet-search-results

Appendix

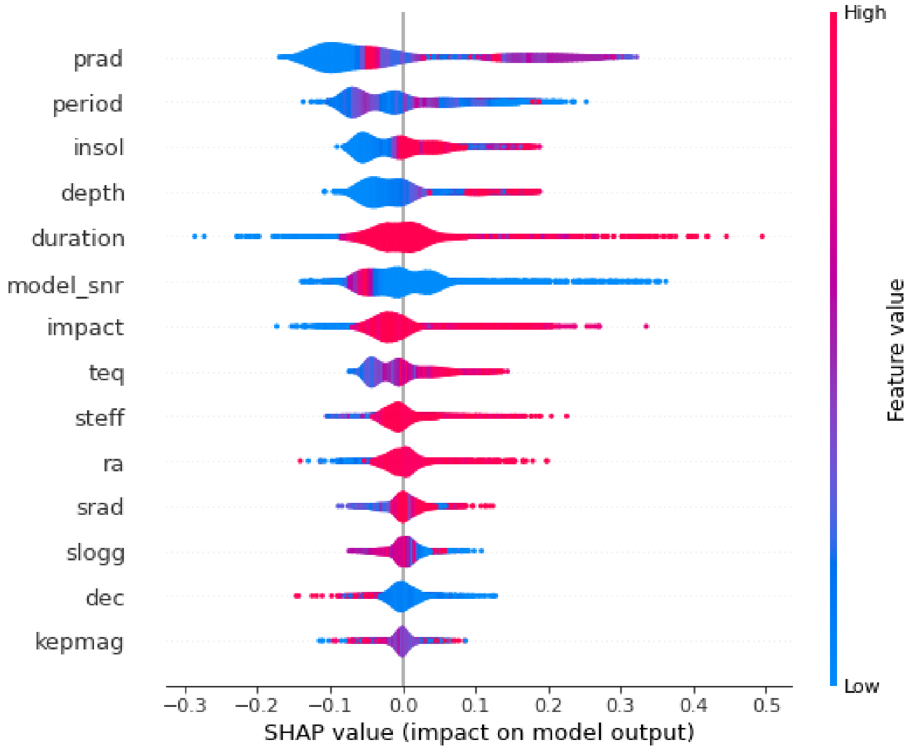

Figure 1: SHAP Feature Importance Violin Plot

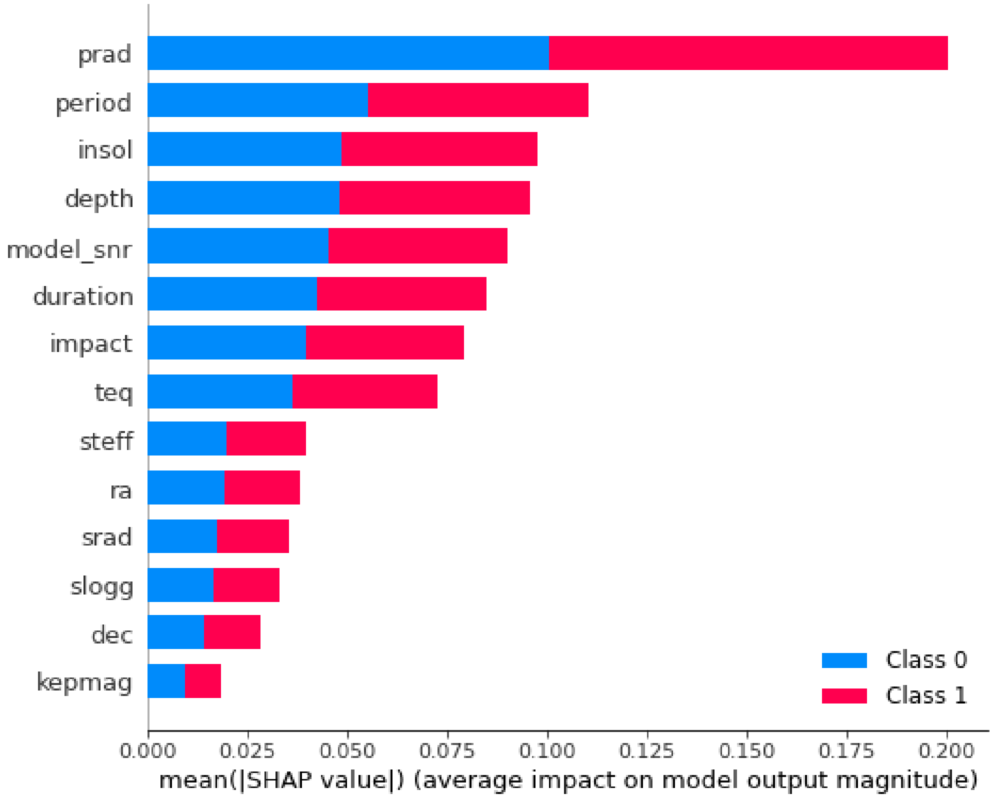

Figure 2: SHAP Feature Importance by Class

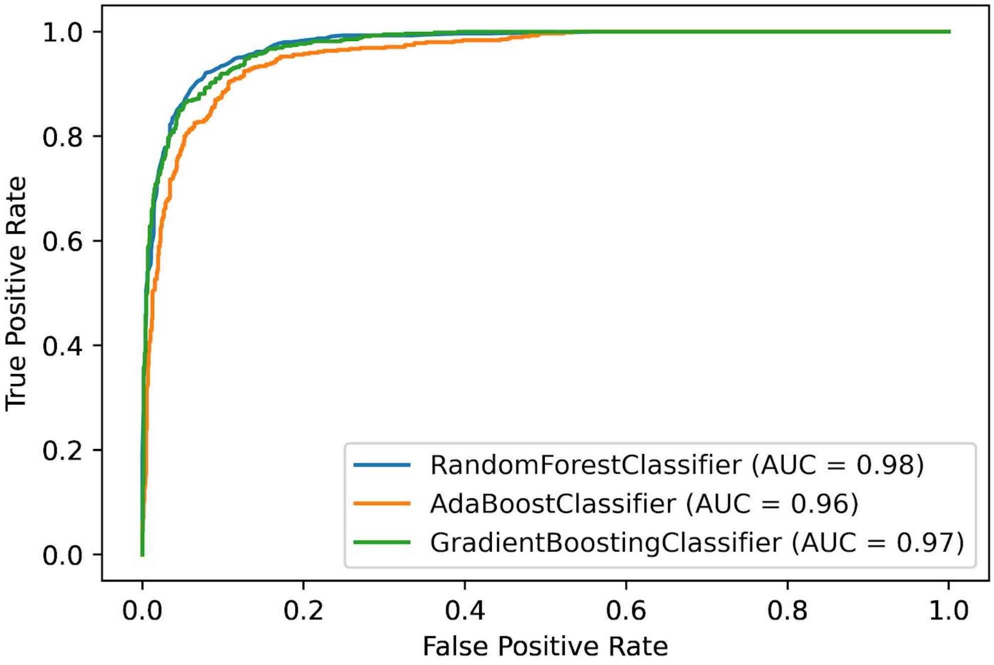

Figure 3: Ensemble Classifier AUC plots

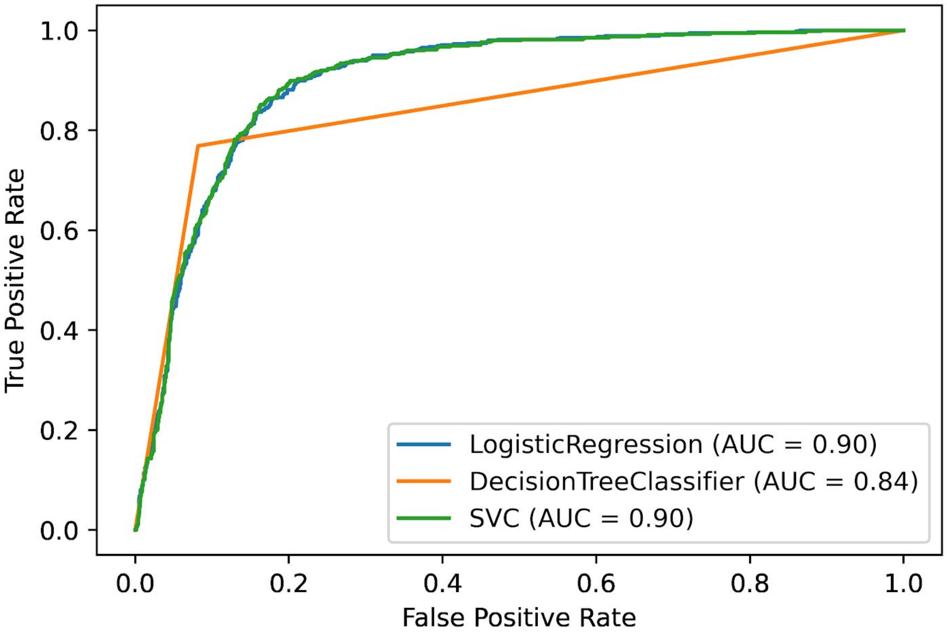

Figure 4. Base model AUC plots

Team Contributions:

Anthony Chen

- Assisted with Python code for preparation, modeling, and visualization
- Write up of Data Preparation section
- Assisted with write up of deployment section
- Edited paper

Sharder Islam

- Assisted with Python code for preparation, modeling, and visualization
- Write up of Modeling & Evaluation section
- Assisted with write up of deployment section
- Edited paper

David May

- Assisted with Python code for preparation, modeling, and visualization
- SHAP plotting
- Write up of Data Understanding section
- Assisted with write up of deployment section
- Edited paper

Daniel Tang

- Assisted with Python code for preparation, modeling, and visualization
- Write up of Business Understanding section
- Assisted with write up of deployment section
- Edited paper