ESTIMATING EARNINGS AFTER GRADUATION FROM POST-SECONDARY INSTITUTIONS

Alex Van Rooy

ABSTRACT

- The decision of which post-secondary institution to attend can be a very difficult decision to make.
- Students often want to ensure that the school they are picking will be the best for meeting their future career goals, and help them reach success in their chosen field.
- One common metric of success is earnings.
- Using the machine learning models: SVM, Decision Tree, KNN, and Lasso, create a regression model that uses characteristics from an institution to provide an estimation of what future earnings can look like after graduation from that institution.

- The purpose is to provide students with a tool that can assist them in their search for the institution that best meets their needs, as well as provide some insight into what contributes to earnings post-graduation.

AGENDA

Introduction

Methodology

Results

Conclusion

INTRODUCTION

The Problem:

- Students that plan on pursuing post-secondary education are faced with a selection of different schools that they could attend.
- For some students the choice is easy, but for many the choice of institution can be overwhelming.
- Students want to pick a school that offers them the best chance at being successful in their future careers, one large metric to measure success is earnings.

The Solution:

- Through a process of supervised learning, create a machine learning regression model that takes information about a given institution and returns an estimate for future earnings after graduation.

The Goal:

- Design a tool that can help students when assessing potential post-secondary institutions, as well as provide deeper insight into what factors may contribute to earning more after graduation.

METHODOLOGY

THE DATA

- The data was taken from the College Scorecard website which provides a dataset of American colleges and university along with different features of each school.
- The raw dataset has over 6000 samples of data, each with almost 3000 features.
- Information in the dataset was related to topics such as costs, student body demographics, admission rates, earnings, student debt, and more.

CLEANING THE DATA

- The dataset contained a lot of missing values in the form of PrivacySupressed entries.
- The first step in cleaning was finding all the rows and columns that had over 40% of their data listed as NaN and removing them.
- After this procedure, the dataset had about 5000 rows and about 700 columns that remained.
- For the remaining NaN values, the missing values were replaced with the column mean.

FEATURE SELECTION

- The number of features was over 700, this is far too many.
- Due to processing constraints, assessing the importance of all 700 features would not have been feasible, instead around 80 features were hand selected based off relevance to the goal of the project.
- For each institution in the dataset, there was a feature that represented the mean earnings of a student, 6 years after graduation. This feature was used as the dependent variable.
- Once the independent variables and the dependent variable were selected, the data was split into training and testing sets using a 70/30 split.

TRAINING THE MODELS

- The machine learning models used in the project were:
  - SVM
  - Decision Tree
  - KNN
  - Lasso

- Each model followed a similar training procedure. The hyperparameters were identified for each model and tuned until acceptable results were achieved.

- For models that were overfitting to the training data, bagging was used as a regularization technique.

- After a series of training sessions, the parameters that performed best for each model were chosen to build the final model that will be tested on the test data.

RESULTS

RESULTS ANALYSIS

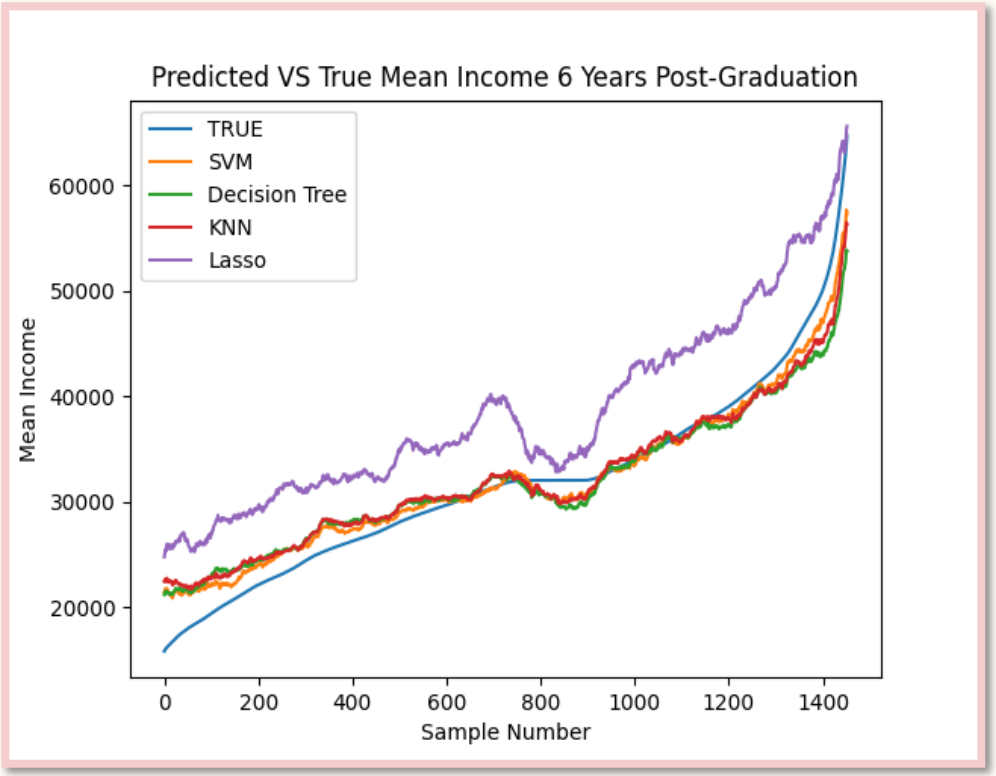
- The fitted models were tested using the same, unseen, testing set.
- After each model produced predictions for the test data, the model's predicted values were compared with the true values.
- The performance of each model was evaluating using the following metrics:
  - R-Squared
  - Root Mean Square Error (RMSE)
  - Mean Absolute Error (MAE)

RESULTS TABLE

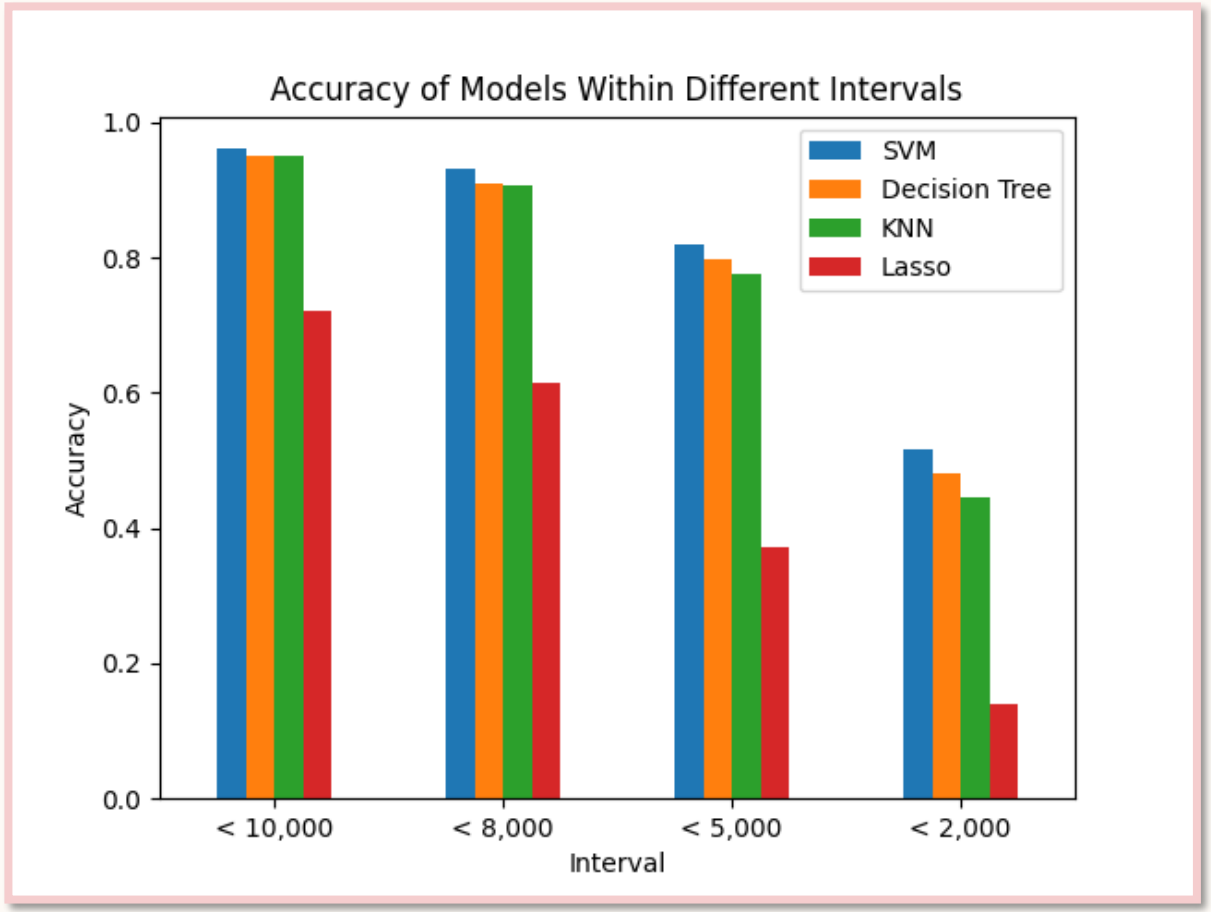
Model	R-Squared	Root Mean Square Error	Mean Absolute Error
SVM	0.8	4527.58	2965.03
Decision Tree	0.77	4833.70	3232.26
KNN	0.77	4911.69	3356.71
Lasso	0.09	9709.69	7705.84

RESULTS ANALYSIS

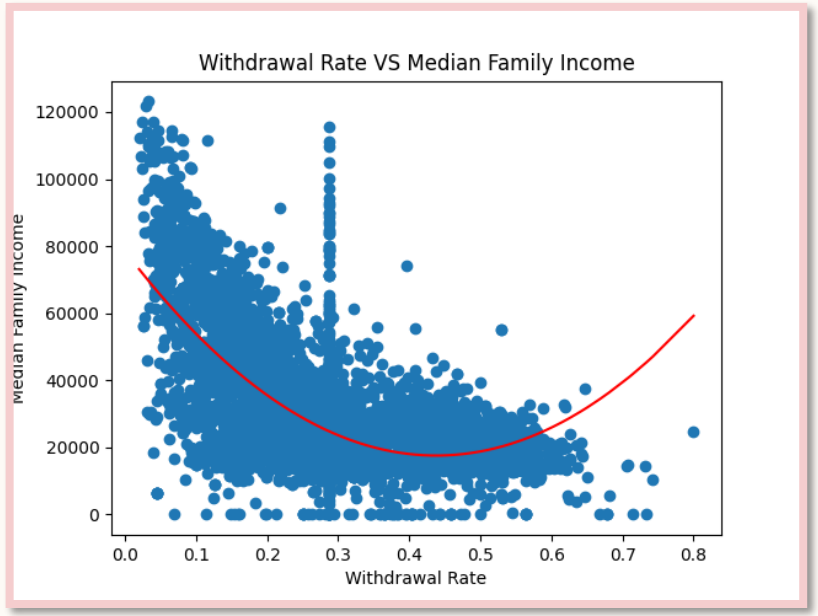
- SVM outperformed all models across all metrics.
- The worst performing model was the Lasso model, however this was expected as the nature of the data was non-linear and Lasso is a linear model.
- In general, the more flexible models performed best on the data.
- The models all seemed to follow a similar shape, this could mean that there are certain factors that are being overlooked in the training process.

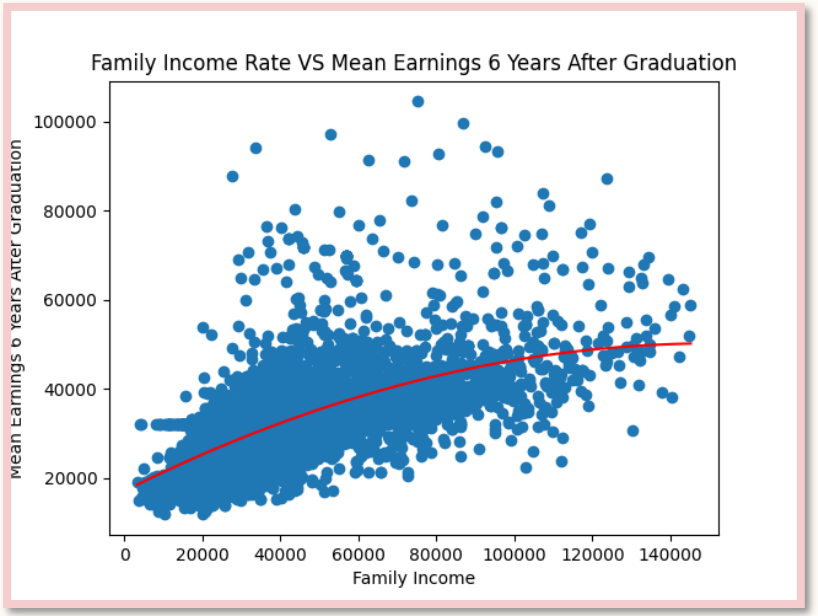


MODEL ACCURACY



CORRELATIONS





- These two figures highlight an important reality that is faced by many students.
- The existing wealth of a student's family does play a big role in the prospective success of that student during their studies and even after they graduate.

CONCLUSION

- A model was created that is able to provide an estimate of future earnings given information on a particular school, with an error of about $4,500.
- The project was able to develop a model that fit data and provide a meaningful result to the end user.
- Students can use this model to evaluate schools they are considering based on how much they may earn 6 years after graduation.

- Future work would involve creating a model that is more interpretable. The model used in this project provided good results, but for a student who is trying to understand what effects the estimation, it may be difficult to find connections.
- Another area of future work would be better pre-processing of the data. Judging by the shape of the models results and the metrics used for each model, it is clear that there may be some factors that are not being accounted for.

THANK YOU