The 2014 *XGBoost: A Scalable Tree Boosting System* paper by Tianqi Chen and Carlos Guestrin describes a new, novel method for scalable tree boosting and it stands for Extreme Gradient Boosting. The main advantages this paper proposes is performance without sacrificing scalability and a package written in multiple languages that is open source from the start. XGBoost is one of the most important algorithms in machine learning considering the various successes it has boasted in recent years, including in the Netflix challenge. The main algorithm behind the XGBoost library is the gradient boosting decision tree. It uses the idea of boosting – an ensemble learning technique – to account for errors that weak trees might have missed after being built. XGBoost includes a couple of important features, including regularization learning, gradient tree boosting, a shrinkage parameter, and column subsampling.

The Census Income dataset is based on the 1994 Census with most of the credit being attributed to Barry Becker. The main task of the dataset is to classify whether a particular person will make more than $50,000 and a variety of predictors from the 1994 census are given to aid in that decision. The predictors included in the base dataset include a person's age, his/her class of work, the education level he/she has attained, his/her marital status, his/her occupation, his/her relationship, his/her race, his/her gender, his/her capital gain/loss for that particular year, the number of hours he/she works in a week, and finally his/her native country.

The primary goal of this project is to compare the performance, in terms of a variety of metrics related to classification tasks, and speed of different machine learning algorithms. Primarily, the base algorithm for comparison would be XGBoost and looking into what separates the XGBoost algorithm from other similar and non-similar – such as a logistic regression – methods. Finally, it would be of interest to know which particular features had the most important impact on the performance of the XGBoost algorithm, so variable importance will also be tracked.