Final Report: Credit One Lending Analysis

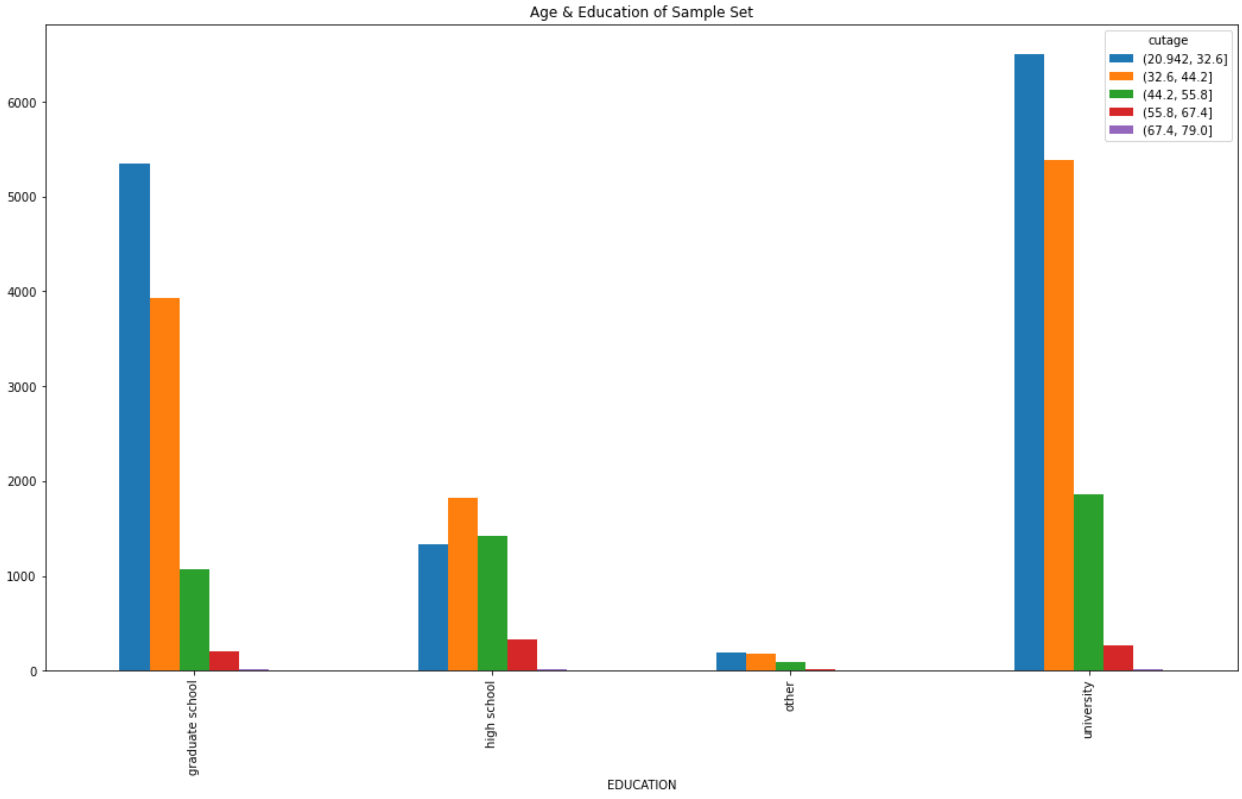
1. **Cleaning and pre-processing Data**

   The dataset used in this analysis was grabbed from a data server using an SQL query. The data that was imported came with some flaws, such as feature headings as the first row of data, instances of missing data segments and repeated instances of observations.

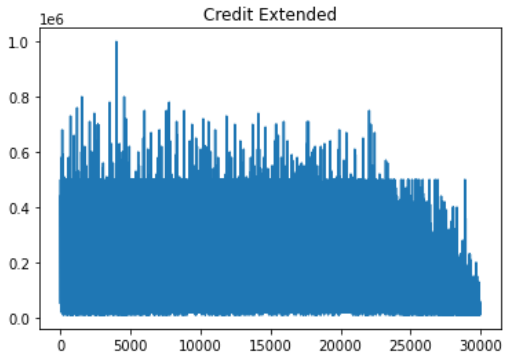
   The solution was to reorganize the data neatly so that only single observations with complete data were included. This reduced the sample size from 30,204 to 29,965 (less than 1% total data loss).

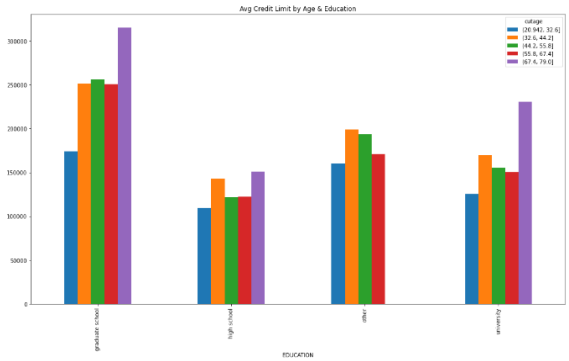
2. **Exploratory Data Analysis**

   Through exploration of the dataset it was discovered the majority of borrowers held a university education or higher. Splitting the ages of borrowers into 5 equal sized bins revealed the following population makeup:

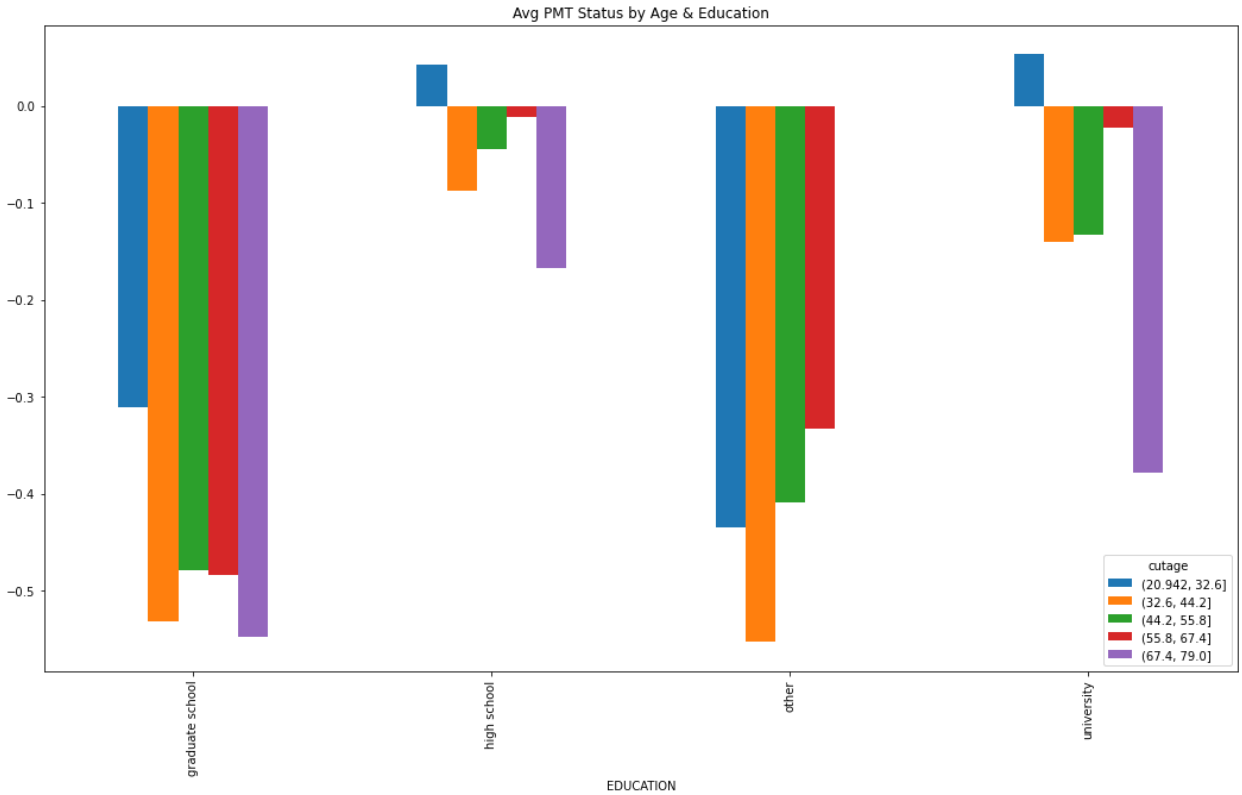


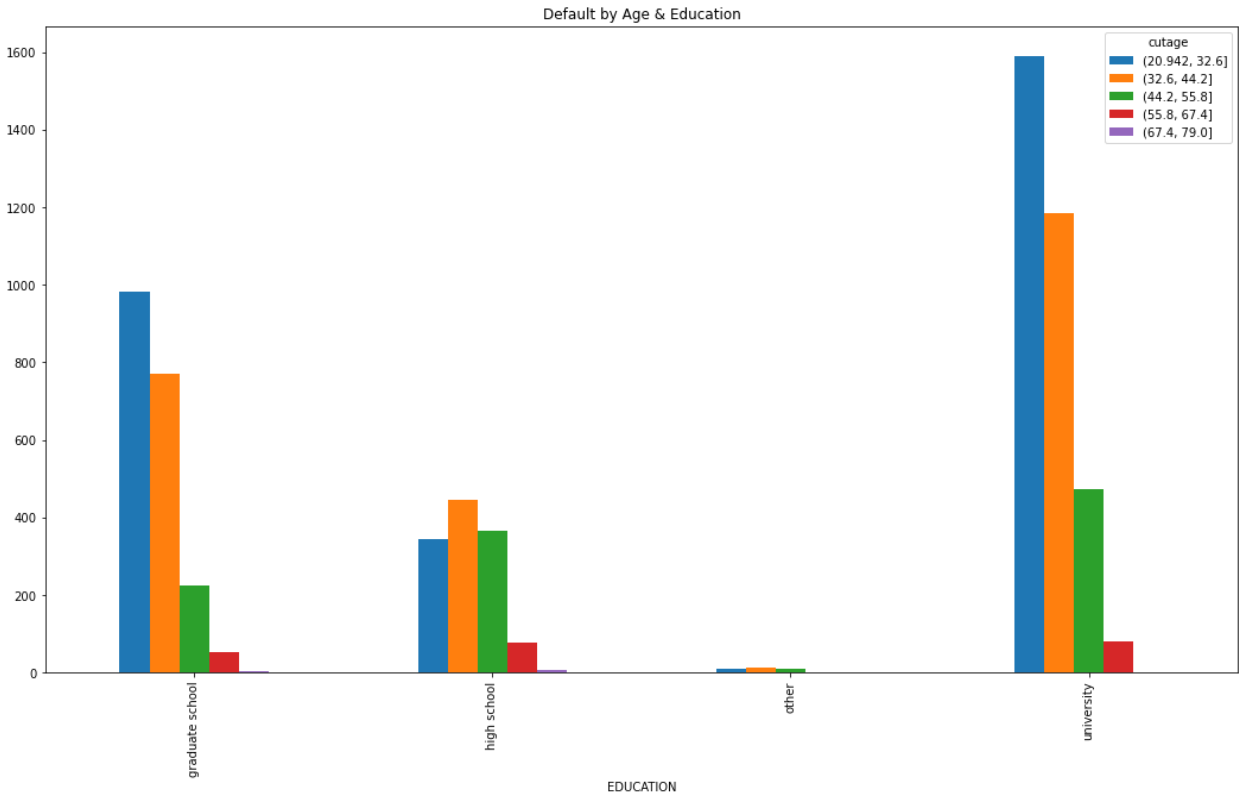
   Despite the diversity of ages and education levels, the total credit loaned out to borrowers appeared more or less banded. This indicates a lack of robust inputs towards determining total credit limit extended since most people end up with the same credit line:





The average payment status was calculated across age and education levels as well. It would appear that most age/education groups average somewhere below “0” which indicates they are either paying their debts in full or utilizing the debt revolver. The youngest of borrowers with university education or high school education averages somewhere between using their credit revolver or falling behind on payments. Of these two groups, defaults are highest in the younger age groups who have achieved university education:





3. Model Building

Three models were chosen for use in this task: Random Forest Classifier, Gradient Boosting Classifier, and Random Forest Classifier. These models split the sample data 70/30 into training and testing data (chosen randomly) to optimize the model. Here are the results of each models predictive accuracy showing GBC to be the best of algorithms.

```
Random Forest Classifier accuracy score is 0.8174015074384351
Decision Tree Classifier accuracy score is 0.7327770602877163
Gradient Boosting Classifier accuracy score is 0.8233132163306194
```

4. Model Evaluation & Results

With Gradient boosting classifier identified as the superior classification algorithm to determine likelihood of default, the following results were achieved:

	precision	recall	f1-score	support
0	0.66	0.37	0.47	1994
1	0.84	0.95	0.89	6996
accuracy			0.82	8990
macro avg	0.75	0.66	0.68	8990
weighted avg	0.80	0.82	0.80	8990

With an accuracy score of 82%, this model can better predict borrowers who won't default (the business objective) compared to the older method which resulted in a default rate of 22%.