



# Predicting Repayment of Education Loans

NU Big Data & Data Science Diploma – Final Project

## Abstract

This document describes the typical phases of a project, the tasks involved with each phase, and an explanation of the outputs of these tasks.



## Contents

1.	Project Team.....	2
2.	Implementation Model & Methodology .....	2
3.	Business Understanding .....	3
3.1	Describing Problem Area .....	3
3.2	Defining Business Objectives .....	3
3.3	Business Success Criteria .....	4
3.4	Data Sources & Knowledge Stores .....	4
3.5	Project Plan.....	4
3.6	Assessing Tools & Techniques .....	4
4.	Data Understanding.....	5
4.1	Describing Data.....	5
4.2	Exploring Data.....	5
4.3	Verifying Data Quality.....	11
5.	Data Preparation.....	11
5.1	Selecting Data (Including or Excluding Data).....	11
5.2	Cleaning Data .....	20
5.3	Constructing New Data.....	21
6.	Modeling .....	21
6.1	Selecting Modeling Techniques.....	21
6.2	Generating a Test Design.....	21
6.3	Building the Models.....	21
6.3.1	Parameter Settings .....	21
6.3.2	Running the Models.....	21
6.4	Assessing the Model .....	21
7.	Evaluation .....	23
7.1	Evaluating the Results.....	23
8.	Deployment .....	23
8.1	System Main Components.....	23
8.2	System Workflow.....	24
9.	References .....	25



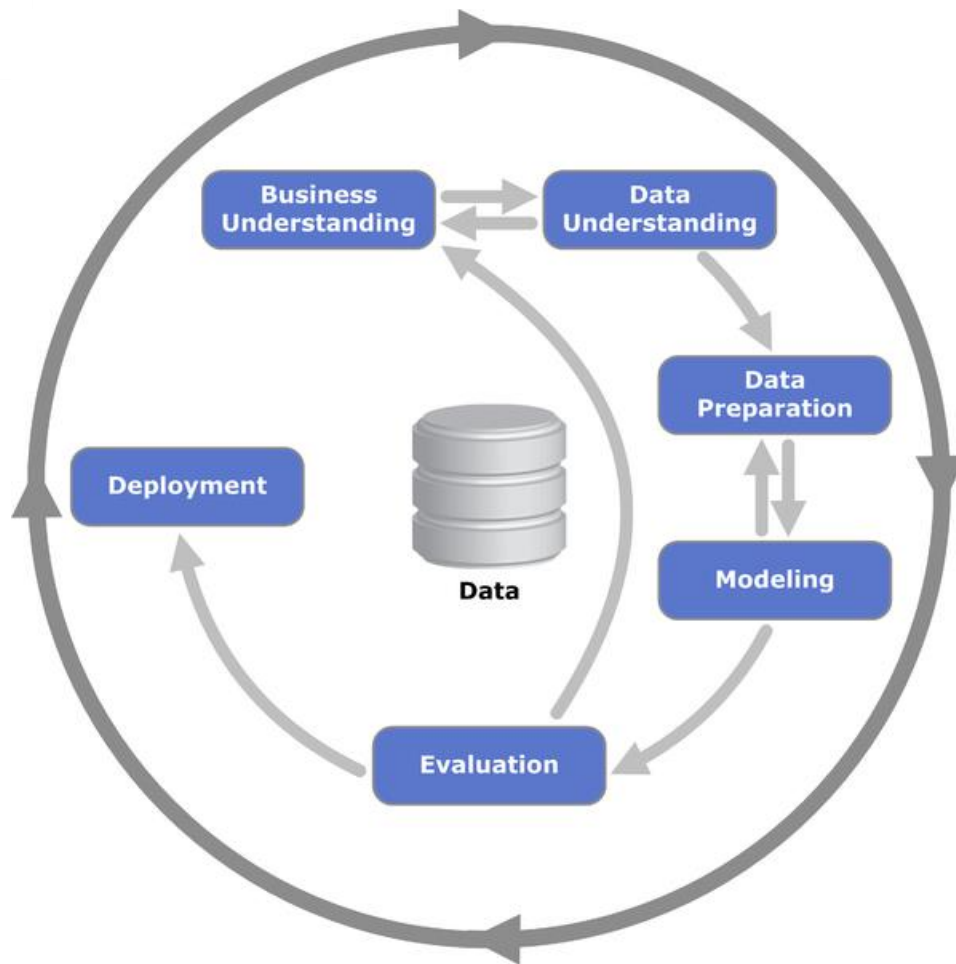
## 1. Project Team

“Miners” Team:

1. Heba Saad
2. HebatAllah Zakaria
3. Usama Atteya
4. Yehia Atef Dorgham

## 2. Implementation Model & Methodology

- This report presents our work on building a model that predicts the students’ loan repayment rates.
- The dependent variable is present in the dataset in a disaggregated form and hence it is to be decided which one of them is to be used. These are continuous variables that contain many unidentified values. Our model will predict those values with good accuracy.
- This document describes the entire process in the form of CRISP-DM model which represents the standard life cycle model for any data mining project.
- The life cycle model consists of six phases as shown in the following diagram, with arrows indicating the most important and frequent dependencies between phases.



### 3. Business Understanding

#### 3.1 Describing Problem Area

- Our project's idea about Educational Loans Repayment
- Most students during their college education incur a significant amount of debt
- In an effort to make educational investments less speculative, the US Department of Education has matched information from the student financial aid system with federal tax returns to create the College Scorecard dataset.

#### 3.2 Defining Business Objectives

- Our objective is to use the current institutes' dataset to predict students' ability to repay their educational loans by exploring different institutional features.
- The proposed predictive model will aid the decision-makers of the US government to minimize the risk of bad debts.
- Also, it will put the power in the hands of students and families to compare colleges and see a better vision of how many graduates at a particular school are able to pay back their student loans.

- The ability to pay back student loans is generally a good indicator of how well colleges and universities are preparing their graduates for the job market.

### 3.3 Business Success Criteria

- Minimizing the loss probability due to non-repaid loans for Loan granting organizations.

### 3.4 Data Sources & Knowledge Stores

- We depend on a public dataset, US Department of Education (<https://collegescorecard.ed.gov/data/>)
- College Scorecard data are provided through federal reporting from institutions, data on federal financial aid, and tax information.
- These data provide insights into the performance of schools that receive federal financial aid dollars and the outcomes of the students of those schools.


### 3.5 Project Plan

This is a high-level plan shows the project timeline across the main six phases

Phase	Time
Business Understanding	1 Week
Data Understanding	2 Weeks
Data Preparation	2 Weeks
Modeling	2 Week
Evaluation	1 Week
Deployment	1.5 Week
Documentation	1 Week

### 3.6 Assessing Tools & Techniques

- The tools and platforms, used to implement our system were as following:
  - MySQL Database
  - Sqoop
  - Hadoop ( HDFS )
  - PySpark
  - Jupyter
  - Flask
  - Azkaban (Workflow Scheduler)
  - GitHub repository
- The techniques used to implement our system were as following:
  - Data Exploration & Feature understanding
    - Correlation
    - Statistical graphs
    - Null values detection
    - Outliers detection
  - Data Preparation & Preprocessing

- 
- Handling data encrypted or masked due to privacy issue
  - Handling Null/Nan values detection (Imputation)
  - Categorical variables encoding
  - Models
    - XGBoost
    - Ridge
    - Lasso
    - Multiple Linear Regression
    - SVR
  - Model Evaluation & Assessment
    - R Square
    - RMSE
    - Cross-Validation (K-fold)

## 4. Data Understanding

### 4.1 Describing Data

- After review the data definition guide for this dataset, we found that data can be categorized into eight main categories as following:
  1. Basic information about the dataset (OPEID, Currently Operating...)
  2. About the School data(identifiers, location, degree type and profile, programs offered, and the academic profile of students enrolled)
  3. Academics and Admissions data
  4. Costs data (to evaluate the tradeoffs of access, affordability, and outcomes)
  5. Student data (family income, race, Part/full-time status...)
  6. Financial Aid data (including Pell Grants and federal student loans)
  7. Completion data (College completion is associated with other positive outcomes, like finding a job and successfully repaying student loans)
  8. Earnings and Repayment data

### 4.2 Exploring Data

- Actually this data partitioned as 22 CSV file with **1,731** associated columns (variables/features), So we start to study
- There's only one table in the dataset [Scorecard], It has over a hundred thousand rows in it **124699**.
- Each row corresponds to data on a US school in a given year. The years covered in this data range from 1996 to 2013
- This data is very wide and rich: it has 1731 associated columns (variables/features). One of the most interesting aspects of this data is the earnings information it contains on students once they graduate.

- We start to review data definition guide document in depth to:
  - Understand these 1731 associated features, their type, their business meaning, and the relations between them.
  - Validate the data availability and consistency over the different years, as some features defined at specific years and obsolete at another one due to the changes in laws and business regulations.
- Then as per our initial understanding of the use case and data availability, we choose to:
  - Start with a sample of data (e.g. Data of 2013-2014 academic year).
  - Select 56 features out of these 1731 associated features, those are candidates to affect the target of this use case.
- only suspected features (inputs and outputs) will be loaded to examine by graphs and statistics

Feature	Description
INSTNM	The institution's name
ADM_RATE , ADM_RATE_ALL	<b>Admission Rate</b> For institutions with multiple branches, ADM_RATE includes the admissions rate at each campus, while ADM_RATE_ALL represents the admissions rate across all campuses, defined as the total number of admitted undergraduates across all branches divided by the total number of undergraduates who applied across all branches
ACTCMMID , ACTENMID , ACTMTMID , ACTWRMID	ACT (ACT*MID for CM, EN, MT, and WR) scores
SAT_AVG , SAT_AVG_ALL	Average SAT scores for reading , writing and math
SATMTMID	Midian SAT Math score
UGDS	<b>Number of Undergraduate Students</b> Includes the number of degree/certificate-seeking undergraduates enrolled in the fall
HIGHDEG	<b>Degree Type</b> The highest award level conferred at the institution
CONTROL	Identifies whether the institution's governance structure is public, private nonprofit, or private for-profit
INEXPFTE	Instructional expenditures per FTE student
AVGFAC SAL	The average faculty salary
COSTT4_P , COSTT4_A	<b>Average Cost of Attendance, Tuition and Fees</b> The average annual cost of attendance includes tuition and fees, books and supplies, and living expenses for all full-time, first-time, degree-/certificate-seeking undergraduates who receive Title IV aid. For academic year institutions (COSTT4_A) and for program-year institutions (COSTT4_P)
PCTFLOAN	<b>Percent of Undergraduates Receiving Federal Loans</b> Shows the share of undergraduate students who received

	federal loans in a given year. It can provide important context to figures related to debt, repayment, and non-repayment. This figure may be influenced by the eligibility for federal loans and the extent to which students apply for federal loans, as well as by the cost of the programs
PCTPELL	<b>Percentage of Pell Students</b> Shows the share of undergraduate students who received Pell Grants in a given year. This is an important measure of the access an institution provides to low-income students. However, it may not capture all low-income students
MEDIAN_HH_INC	Median household income
UGDS_WHITE , UGDS_BLACK , UGDS_HISP , UGDS_ASIAN , UGDS_AIAN , UGDS_NHPI , UGDS_2MOR , UGDS_NRA , UGDS_UNKN	<b>Undergraduate Student Body by Race and Gender</b> This includes the total enrollment of undergraduate, degree-seeking students, based on fall enrollment, who are: men (UGDS_MEN), women (UGDS_WOMEN), white (UGDS_WHITE), black (UGDS_BLACK), Hispanic (UGDS_HISP), Asian (UGDS_ASIAN), American Indian/Alaska Native (UGDS_AIAN), Native Hawaiian/Pacific Islander (UGDS_NHPI), two or more races (UGDS_2MOR), non-resident aliens (UGDS_NRA), and race unknown (UGDS_UNKN).
PPTUG_EF	<b>Undergraduate Students by Part-Time/Full-Time Status</b> Includes the proportion of degree/certificate-seeking undergraduates enrolled part time in the fall term
TUITIONFEE_IN , TUITIONFEE_OUT , TUITIONFEE_PROG	The cost data include the tuition and required fees of the institution. They are provided for in-state students (TUITIONFEE_IN), out-of-state students (TUITIONFEE_OUT), and program-year institutions (TUITIONFEE_PROG).
TUITFTE	The net tuition revenue per full-time equivalent (FTE) student
DEATH_YR3_RT , DEATH_YR4_RT , COMP_ORIG_YR2_RT , COMP_ORIG_YR3_RT , COMP_ORIG_YR4_RT , LOAN_DEATH_YR3_RT , LOAN_COMP_ORIG_YR3_RT	<b>Completion and Transfer Rates</b> Each institution has all possible outcomes reported: share of students who died (DEATH_YR*_RT), completed at the original institution (COMP_ORIG_YR*_RT), students who ever received a federal loan at the measured institution (LOAN_*)
AGE_ENTRY	Entry age for students
COUNT_NWNE_P10 , COUNT_WNE_P10	
MN_EARN_WNE_P10 , MD_EARN_WNE_P10	<b>Mean and Median Earnings</b> Mean (MN_EARN_WNE_P*) and median (MD_EARN_WNE_P*) earnings are for the institutional aggregate of all federally aided students who enroll in an institution 10 years after the student enrolls
COMPL_RPY_1YR_RT , COMPL_RPY_3YR_RT ,	<b>Repayment Rate on Federal Student Loans</b> These data are available for all borrowers at the institution, as well as

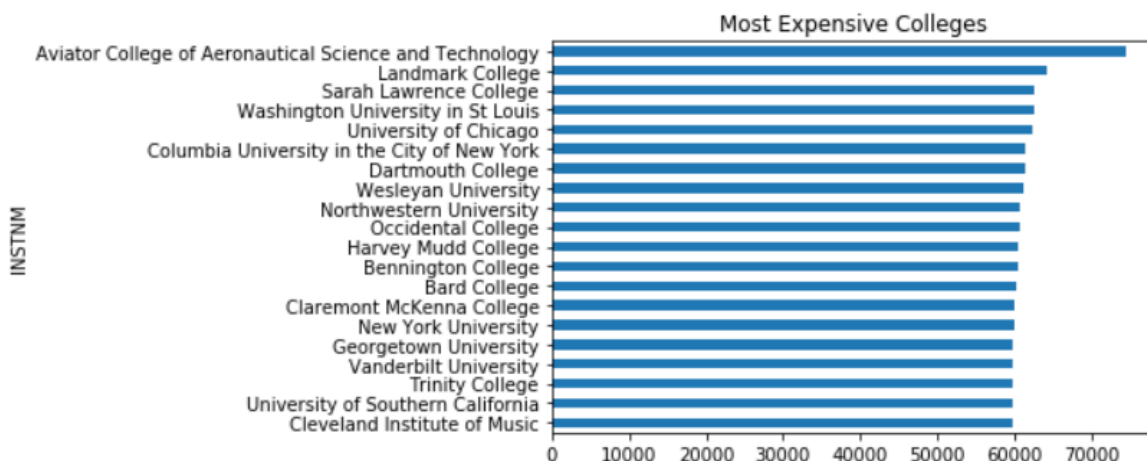


COMPL_RPY_5YR_RT , COMPL_RPY_7YR_RT , NONCOM_RPY_1YR_RT , NONCOM_RPY_3YR_RT , NONCOM_RPY_5YR_RT , NONCOM_RPY_7YR_RT	disaggregated by completion status (COMPL_RPY_* for students who completed and NONCOM_RPY_* for students who withdrew without completing)
------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------

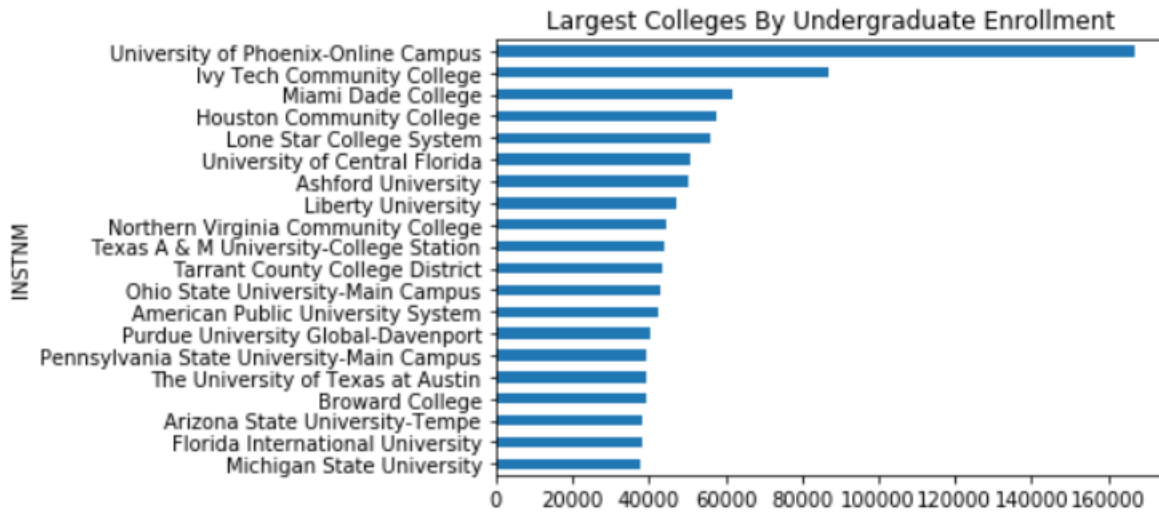
- Explore info of loaded data as following

Index: 7804 entries, Alabama A & M University to Georgia Military College-Stone Mountain  
Data columns (total 53 columns):  
dtypes: float64 (31), int64 (2), object (20)

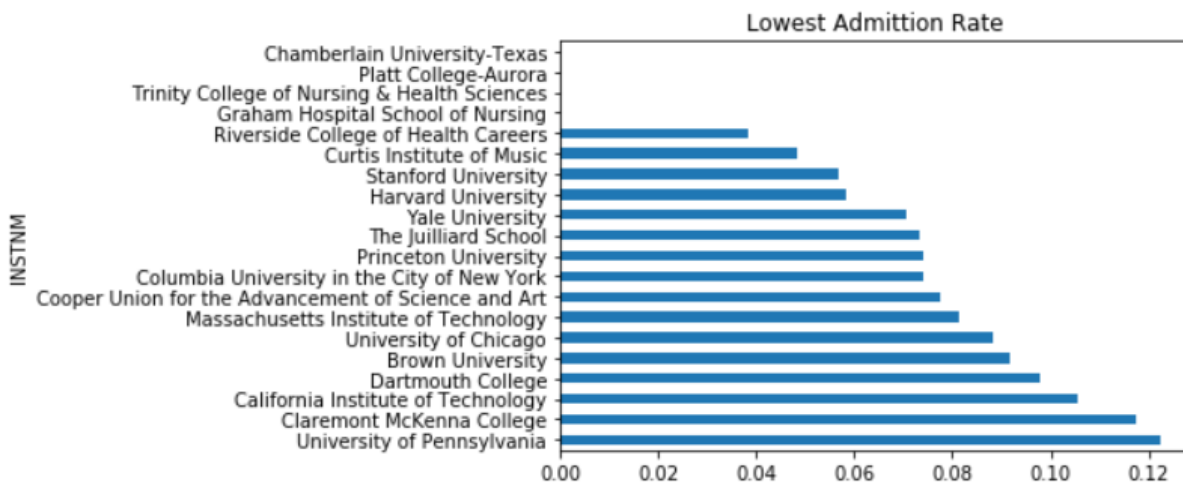
- Produce some graphs show some facts about US colleges.
  - Most expensive colleges
    - The cost here reflects the average annual total cost of attendance, including tuition and fees, books and supplies, and living expenses for all full-time, first-time, degree/certificate-seeking undergraduates who receive Title IV aid.



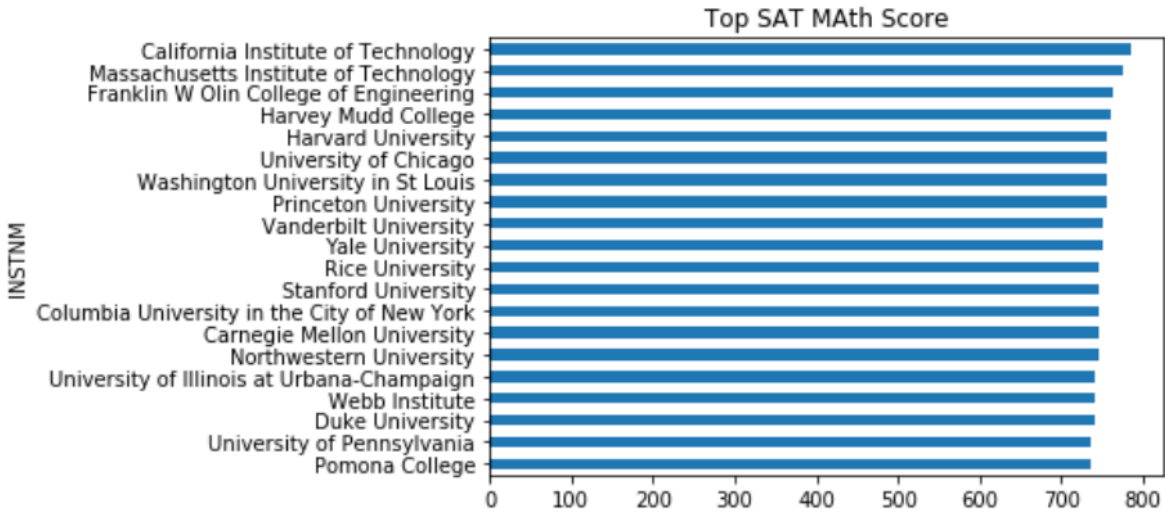
- Colleges have the highest enrollment rate



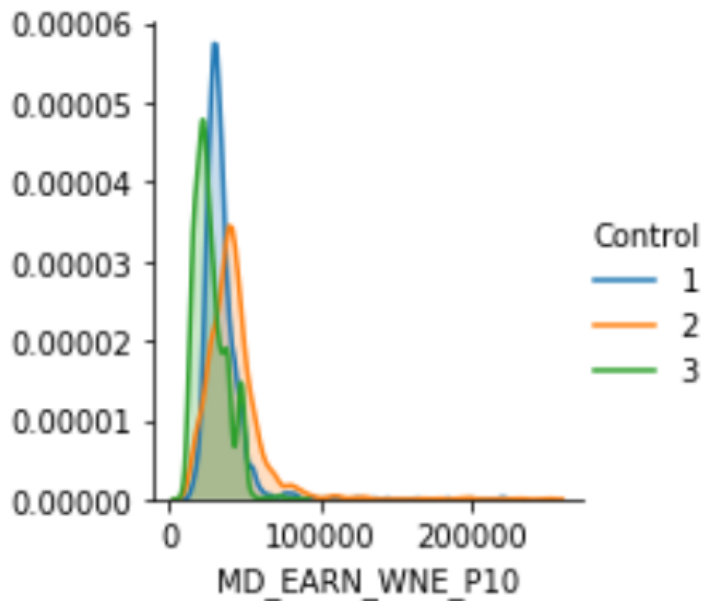
- Colleges have the lowest admission rate
  - One of the most obvious measures of a school's competitiveness is its admission rate.



- Colleges have the top median SAT Math scores



- Median earnings 10 years after matriculation as per college type / control
  - We can see the distribution of earnings 10 years after matriculation at undergraduate institutions. We'll split this into three types of schools:
    - ✓ private non-profit [3],
    - ✓ private for-profit [2],
    - ✓ And public [1].
  - The median earnings for public and private non-profit schools look similar over the bulk of the earnings range.
  - At the very low end, there are more private non-profit schools than public schools. Curiously, the median earnings distribution for private for-profit schools is bimodal.



### 4.3 Verifying Data Quality

- Since our problem is a regression problem, almost all suspected input features are float columns. But some columns have a value of "Privacy Suppressed" as an indicator that the value is missing for privacy reasons.

We found that these values must be replaced by null first to be able to convert these columns to float columns to run correlation, plotting functions safely.

```
def cleanPrivacySuppressed (dataFrame)
```

- Also, we calculated the percentage of missing or null values across all candidate features to be able to decide an appropriate way to handle them. For all model except xgboost we replaced null values by median/mean of column.

## 5. Data Preparation

### 5.1 Selecting Data (Including or Excluding Data)

- As mentioned in section 4.2, we choose to start with 56 features out of these 1731 associated features these 56 features were selected by intuition according to our initial understanding of the available data and the business objectives.
- We tried to choose the best input features and target variable based on **Correlation and business use case**, So we generate the absolute correlation for all variables loaded in data-frame
- There is more than one output variable can be used as the target variable, suspected outputs are :
  - COMPL\_RPY\_1YR\_RT (One-year repayment rate for completers)
  - COMPL\_RPY\_3YR\_RT (Three-year repayment rate for completers)
  - COMPL\_RPY\_7YR\_RT (Seven-year repayment rate for completers)
  - COMPL\_RPY\_5YR\_RT (Five-year repayment rate for completers)
  - NONCOM\_RPY\_3YR\_RT (Three-year repayment rate for non-completers)
  - NONCOM\_RPY\_5YR\_RT (Five-year repayment rate for non-completers)
  - NONCOM\_RPY\_1YR\_RT (One-year repayment rate for non-completers)
  - NONCOM\_RPY\_7YR\_RT (Seven-year repayment rate for non-completers)
- The output variable that has the highest correlation with the other input features will be chosen as the target variable, meanwhile, other candidates will not be considered as input features.

Assuming that the accepted correlation value between the output variable an any of input variables must be greater than 0.5

- For [COMPL\_RPY\_1YR\_RT]

COMPL_RPY_3YR_RT	0.946264
NONCOM_RPY_3YR_RT	0.927266
NONCOM_RPY_5YR_RT	0.913966
NONCOM_RPY_1YR_RT	0.906497
COMPL_RPY_5YR_RT	0.901021




NONCOM_RPY_7YR_RT	0.840605
COMPL_RPY_7YR_RT	0.749900
PCTPELL	0.697547
SAT_AVG_ALL	0.668902
SAT_AVG	0.651445
ACTCMMID	0.636761
ACTMTMID	0.628290
SATMTMID	0.626058
AVGFACSAL	0.607634
ACTENMID	0.605319
MD_EARN_WNE_P10	0.591054
MN_EARN_WNE_P10	0.590682
AGE_ENTRY	0.581915
CONTROL	0.547110
LOAN_DEATH_YR3_RT	0.536322

○ For [COMPL\_RPY\_3YR\_RT]

COMPL_RPY_5YR_RT	0.952187
COMPL_RPY_1YR_RT	0.946264
NONCOM_RPY_5YR_RT	0.928185
NONCOM_RPY_3YR_RT	0.907603
NONCOM_RPY_7YR_RT	0.898931
NONCOM_RPY_1YR_RT	0.865118
COMPL_RPY_7YR_RT	0.851590
LOAN_DEATH_YR3_RT	0.707518
PCTPELL	0.704221
MD_EARN_WNE_P10	0.645223
SAT_AVG_ALL	0.641538
MN_EARN_WNE_P10	0.640586
SAT_AVG	0.623695
AVGFACSAL	0.615004
ACTCMMID	0.614424
ACTMTMID	0.606382
SATMTMID	0.591391
CONTROL	0.585123
ACTENMID	0.582368
AGE_ENTRY	0.550468
HIGHDEG	0.535670
COMP_ORIG_YR2_RT	0.507860

○ For [COMPL\_RPY\_5YR\_RT]

COMPL_RPY_3YR_RT	0.952187
NONCOM_RPY_7YR_RT	0.916724
NONCOM_RPY_5YR_RT	0.908556
COMPL_RPY_1YR_RT	0.901021
COMPL_RPY_7YR_RT	0.890355
NONCOM_RPY_3YR_RT	0.871874
NONCOM_RPY_1YR_RT	0.820319
PCTPELL	0.686690
LOAN_DEATH_YR3_RT	0.659897
MD_EARN_WNE_P10	0.645418
MN_EARN_WNE_P10	0.640329



SAT_AVG_ALL	0.625157
AVGFACSAL	0.624277
SAT_AVG	0.618989
ACTCMMID	0.613180
ACTMTMID	0.605468
SATMTMID	0.593424
ACTENMID	0.581160
CONTROL	0.565434
AGE_ENTRY	0.554506
COMP_ORIG_YR2_RT	0.525298
HIGHDEG	0.518821

○ For [COMPL\_RPY\_7YR\_RT]

COMPL_RPY_5YR_RT	0.890355
NONCOM_RPY_7YR_RT	0.873590
COMPL_RPY_3YR_RT	0.851590
NONCOM_RPY_5YR_RT	0.803168
COMPL_RPY_1YR_RT	0.749900
NONCOM_RPY_3YR_RT	0.746289
LOAN_DEATH_YR3_RT	0.712669
MD_EARN_WNE_P10	0.707403
MN_EARN_WNE_P10	0.696691
NONCOM_RPY_1YR_RT	0.667975
COMP_ORIG_YR2_RT	0.634738
PCTPELL	0.578859
AVGFACSAL	0.571819
SAT_AVG_ALL	0.566967
HIGHDEG	0.565509
ACTMTMID	0.548090
ACTCMMID	0.544406
SAT_AVG	0.543223
SATMTMID	0.521642
ACTENMID	0.500940

○ For [NONCOM\_RPY\_1YR\_RT]

NONCOM_RPY_3YR_RT	0.944075
COMPL_RPY_1YR_RT	0.906497
NONCOM_RPY_5YR_RT	0.904081
COMPL_RPY_3YR_RT	0.865118
COMPL_RPY_5YR_RT	0.820319
NONCOM_RPY_7YR_RT	0.812678
PCTPELL	0.673040
COMPL_RPY_7YR_RT	0.667975
SAT_AVG	0.662759
SAT_AVG_ALL	0.653900
ACTCMMID	0.646307
ACTMTMID	0.636790
SATMTMID	0.635962
ACTENMID	0.621458
AVGFACSAL	0.588600
MN_EARN_WNE_P10	0.586222
MD_EARN_WNE_P10	0.584202



TUITIONFEE_OUT	0.571291
AGE_ENTRY	0.557894

○ For [NONCOM\_RPY\_3YR\_RT]

NONCOM_RPY_5YR_RT	0.946722
NONCOM_RPY_1YR_RT	0.944075
COMPL_RPY_1YR_RT	0.927266
COMPL_RPY_3YR_RT	0.907603
COMPL_RPY_5YR_RT	0.871874
NONCOM_RPY_7YR_RT	0.869925
COMPL_RPY_7YR_RT	0.746289
PCTPELL	0.711239
SAT_AVG	0.653461
SAT_AVG_ALL	0.647873
ACTCMMID	0.646958
ACTMTMID	0.644148
MD_EARN_WNE_P10	0.629235
AVGFACSAL	0.628608
MN_EARN_WNE_P10	0.625422
ACTENMID	0.623343
SATMTMID	0.620496
TUITIONFEE_OUT	0.569642
AGE_ENTRY	0.562711
CONTROL	0.533233
HIGHDEG	0.509697

○ For [NONCOM\_RPY\_5YR\_RT]

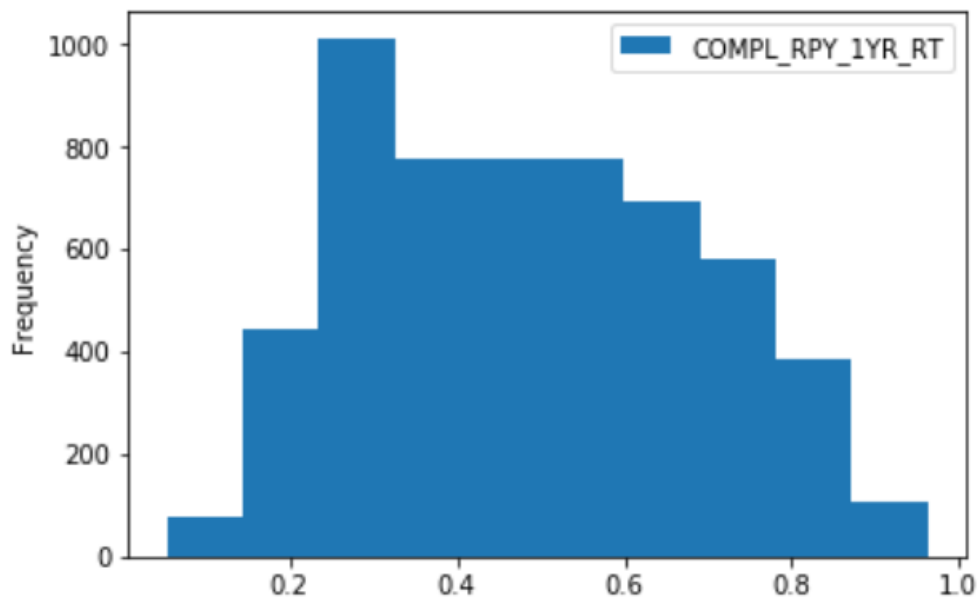
NONCOM_RPY_3YR_RT	0.946722
COMPL_RPY_3YR_RT	0.928185
NONCOM_RPY_7YR_RT	0.923287
COMPL_RPY_1YR_RT	0.913966
COMPL_RPY_5YR_RT	0.908556
NONCOM_RPY_1YR_RT	0.904081
COMPL_RPY_7YR_RT	0.803168
PCTPELL	0.722585
AVGFACSAL	0.652072
SAT_AVG_ALL	0.650532
MD_EARN_WNE_P10	0.650422
SAT_AVG	0.648010
MN_EARN_WNE_P10	0.645042
ACTCMMID	0.638323
ACTMTMID	0.633066
SATMTMID	0.618226
ACTENMID	0.604815
AGE_ENTRY	0.571624
CONTROL	0.557939
TUITIONFEE_OUT	0.550776
HIGHDEG	0.529304

○ For [NONCOM\_RPY\_7YR\_RT]

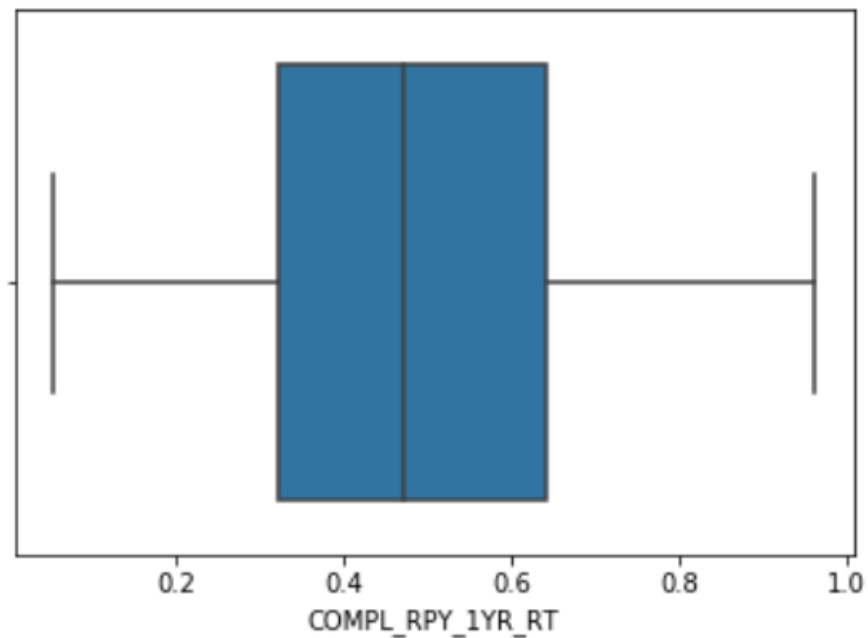


NONCOM_RPY_5YR_RT	0.923287
COMPL_RPY_5YR_RT	0.916724
COMPL_RPY_3YR_RT	0.898931
COMPL_RPY_7YR_RT	0.873590
NONCOM_RPY_3YR_RT	0.869925
COMPL_RPY_1YR_RT	0.840605
NONCOM_RPY_1YR_RT	0.812678
LOAN_DEATH_YR3_RT	0.733568
MD_EARN_WNE_P10	0.691634
MN_EARN_WNE_P10	0.690164
SAT_AVG_ALL	0.683118
ACTMTMID	0.670992
ACTCMMID	0.667276
SAT_AVG	0.667223
AVGFACSAL	0.660287
PCTPELL	0.658074
SATMTMID	0.649815
ACTENMID	0.631715
COMP_ORIG_YR2_RT	0.555561
HIGHDEG	0.553173
TUITIONFEE_OUT	0.552523
CONTROL	0.525132
AGE_ENTRY	0.505059

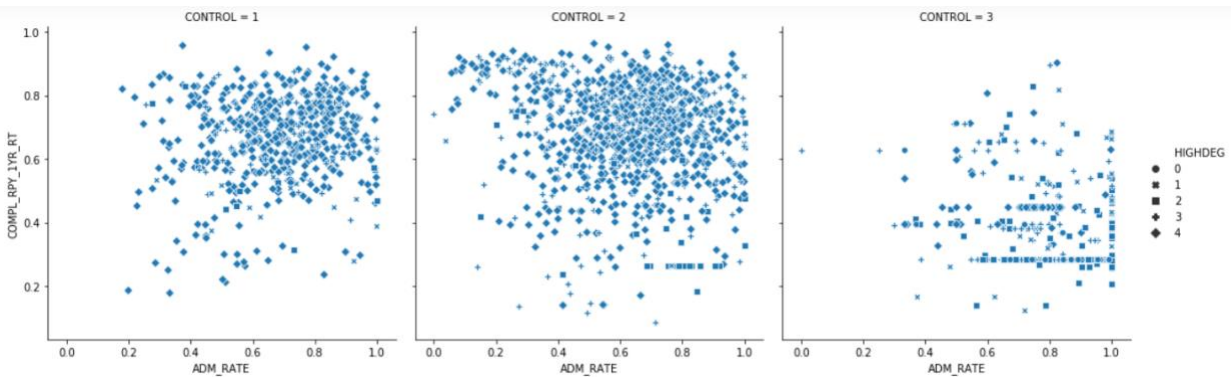
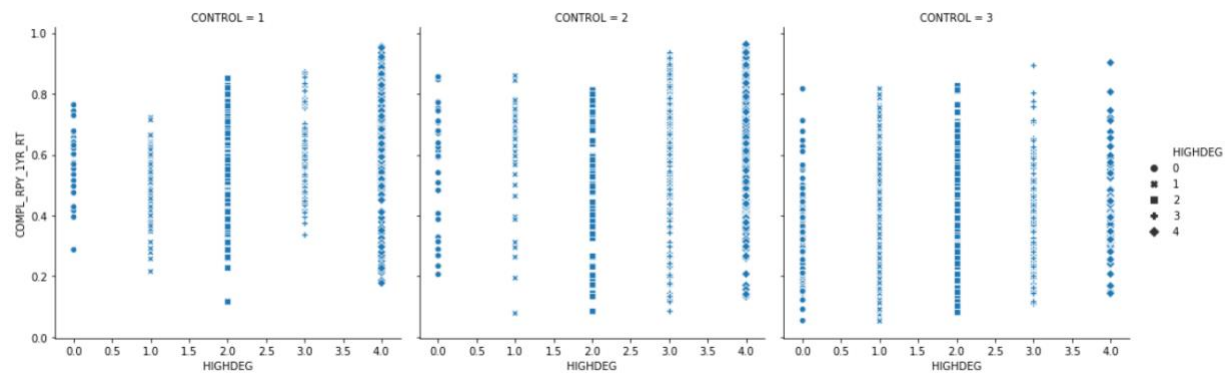
- Since the Correlation of each target with input features is very close to others, **COMPL\_RPY\_1YR\_RT** will be chosen as it is less risky, and drop all other candidates from the loaded dataset.

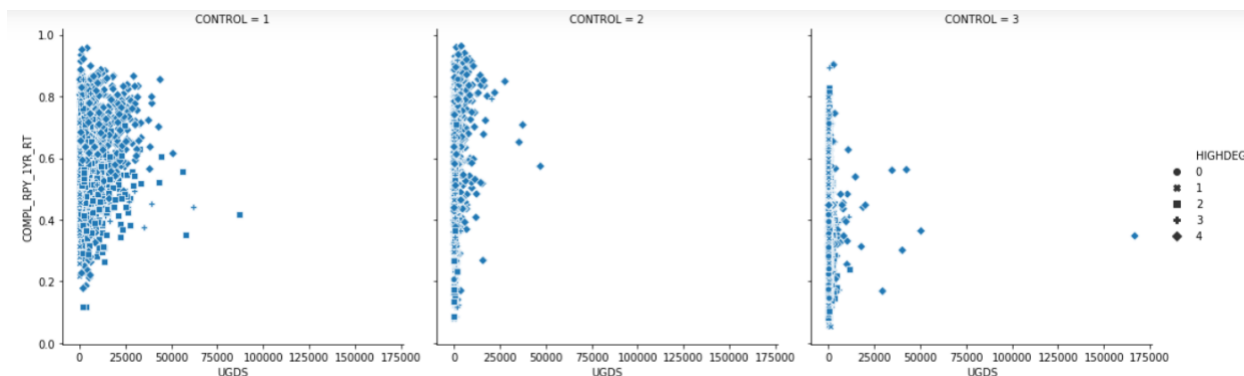
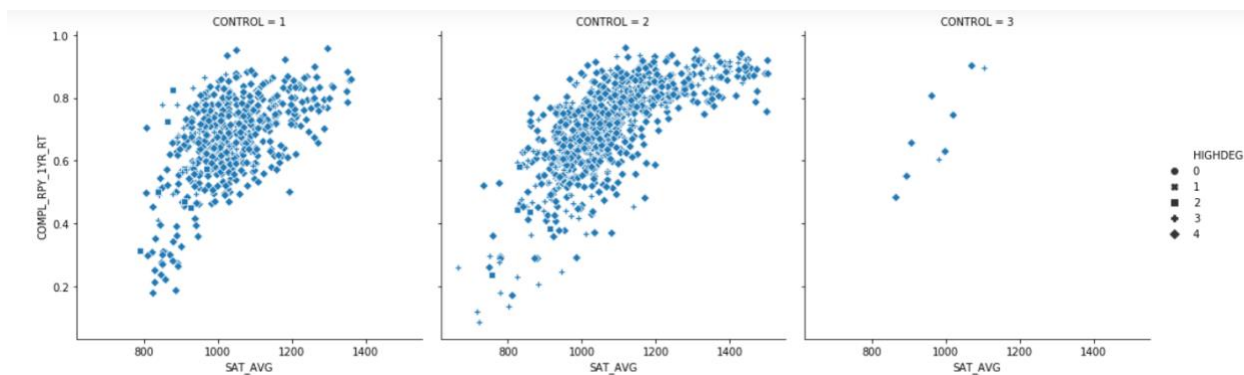
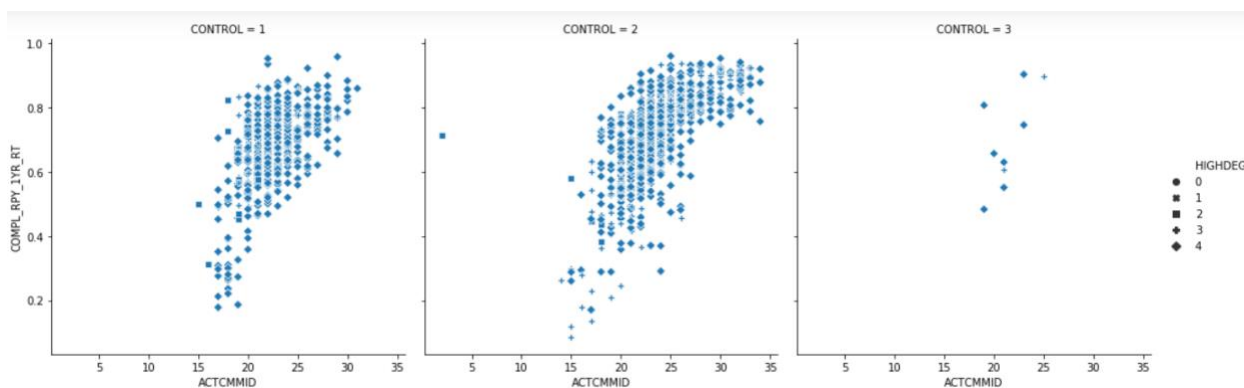
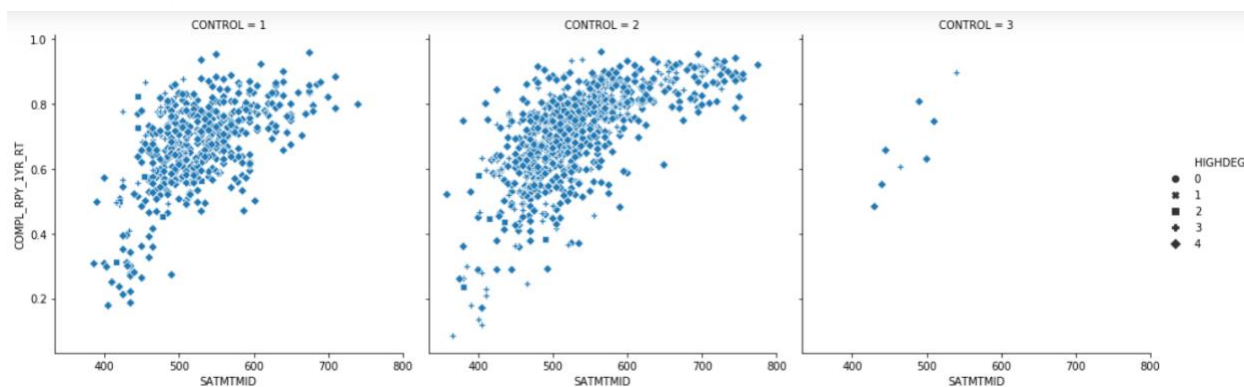




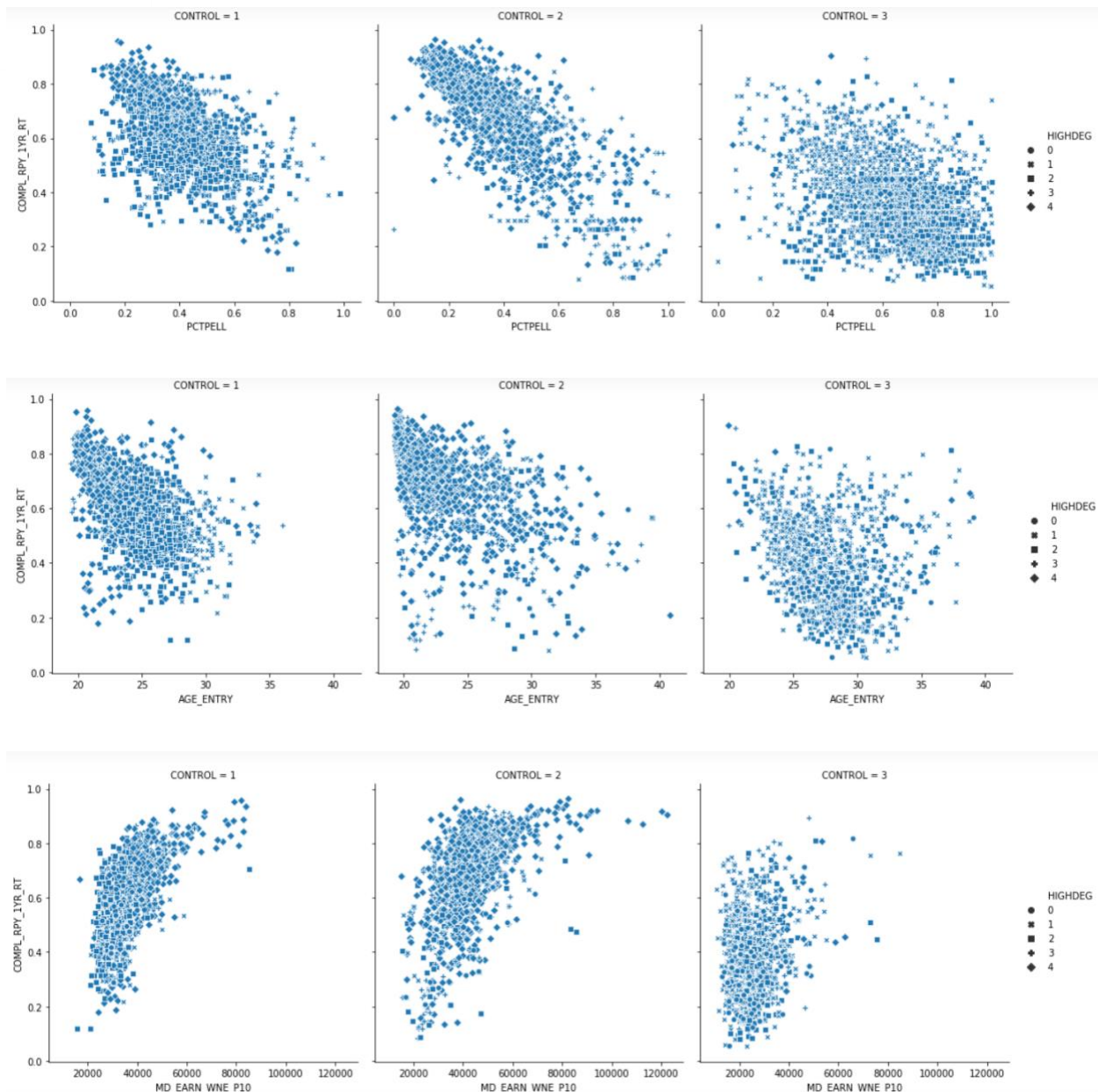


- Then visualize the relationship between the target variable and every input variable to check for any non-linear pattern









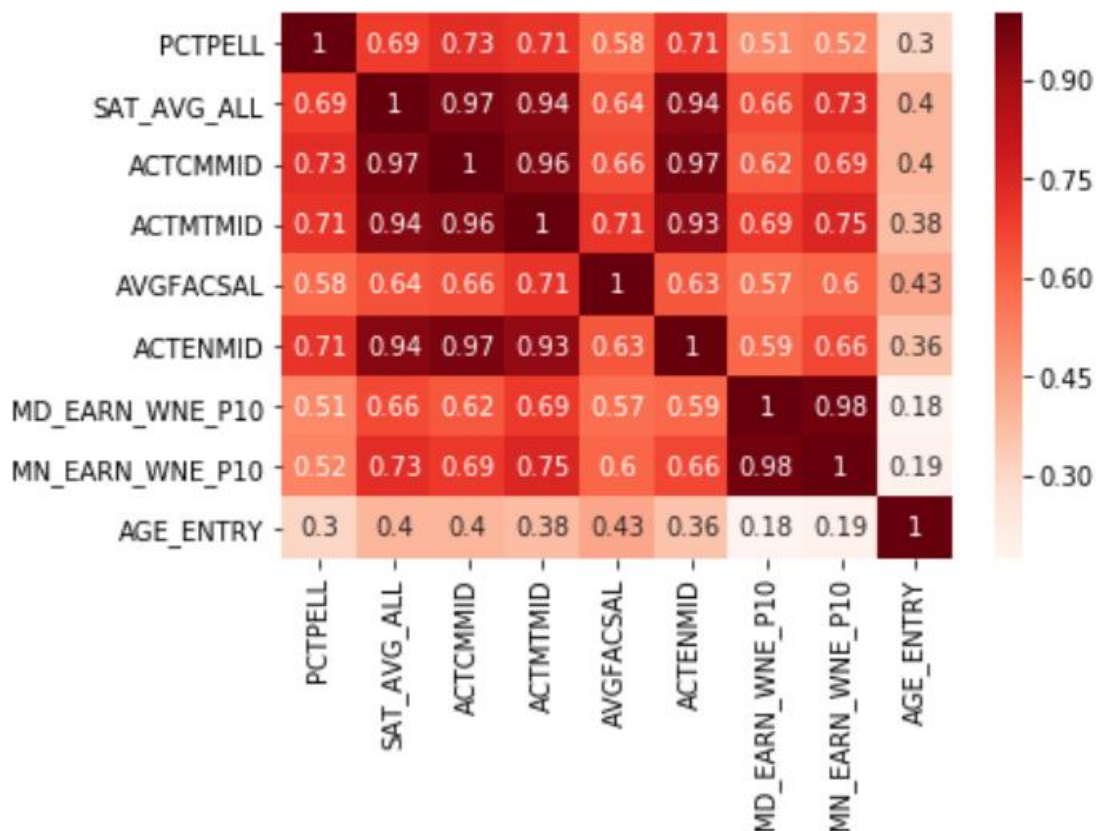
- Input features candidates:

PCTPELL	0.697547
SAT_AVG_ALL	0.668902
ACTCMMID	0.636761
ACTMTMID	0.628290
AVGFACSAL	0.607634



ACTENMID	0.605319
MD_EARN_WNE_P10	0.591054
MN_EARN_WNE_P10	0.590682
AGE_ENTRY	0.581915
CONTROL	0.547110 (not a float column)

- Then we checked the correlation between input features to see the relationship between input features.



## 5.2 Cleaning Data

- Replace all "Privacy Suppressed" values with null value across all features
- Drop all rows that do not have a value for the target variable
- Preprocess the input features, By:
  - Dropping the un-needed columns features (Index **INSTNM** and target output **COMPL\_RPY\_1YR\_RT**)



- Splitting the input features into two types of float features and categorical features as these two types will be treated differently in preprocessing.
- Encode categorical features as a one-hot numeric array, categorical features are **(CONTROL, HIGHDEG)**
- Splitting the data to the training set (80%) and testing set (20%)
- Using SciKit Simple Imputer to replace missing or null values with mean/median values
- Building a pipeline of preprocessing

## 6. Modeling

### 6.1 Selecting Modeling Techniques

- According to the nature of our data and our objective which is predicting the loan repayment probability, we select to start with the following five regression models:
  - ✓ XGBoost
  - ✓ Ridge
  - ✓ Lasso
  - ✓ Multiple Linear Regression
  - ✓ SVR ( with two kernels : Linear , and RBF )
- Since we have a lot of missing values and outliers, it is preferable to start with XGboost. XGboost algorithm is invariant to outliers and can handle missing values by default, so first strategy is replacing any wrong/missing value by null, build the model and check the accuracy.

### 6.2 Generating a Test Design

- The criteria by which the models are assessed is the following:
  - ✓ Calculating MSE, RMSE and r2 values and considering the model that has lower RMSE and higher r2.
  - ✓ Explore feature importance for each model and revising it against the initial findings for correlation and P-Value.

### 6.3 Building the Models


#### 6.3.1 Parameter Settings

- models were built first using their default parameters, then after revising the error and based on the case ( overfitting/ underfitting) we used Grid Search technique to get the best parameters for each model.
- The chosen error function is squared\_error to be sensitive to higher errors.

### 6.4 Assessing the Model

- Chosen evaluation criteria were used to compare different models to choose the best. The table below is the results we got for each model.

	MSE Values		r2 score Values	
	Training Set	Tet Set	Training Set	Test Set
XGBoost	0.0053	0.0075	0.8629	0.8045
Ridge	0.0101	0.0098	0.7388	0.7498
Lasso	0.0210	0.02083	0.4586	0.4716



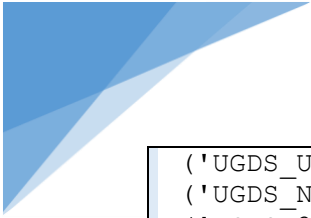
Multiple Linear Regression	0.0100	0.0097	0.7425	0.7525
SVR Linear	0.0354	0.0355	0.0883	0.0985
SVR RBF	0.0079	0.0394	0.7950	-0.0004

- Models error matrix after applying Grid Search and using the best parameters values

	MSE Values		r2 Values	
	Training Set	Tet Set	Training Set	Tet Set
XGBoost	0.003	0.006	0.92	0.834
Ridge	0.0100	0.0098	0.7428	0.7513
Lasso	0.0101	0.0097	0.7401	0.7527
Multiple Linear Regression	0.0100	0.0097	0.7425	0.7525

- XGBoost has been chosen as the best model based on the analysis above. After that we analyzed important. feature importance sorted in descending order for this model is as below:

```
[('CONTROL_Private for-profit', 0.49292025),
 ('SAT_AVG_ALL', 0.2238546),
 ('PCTPELL', 0.05530315),
 ('AGE_ENTRY', 0.033805005),
 ('MN_EARN_WNE_P10', 0.023122836),
 ('MD_EARN_WNE_P10', 0.023000425),
 ('UGDS_BLACK', 0.01871721),
 ('COUNT_NWNE_P10', 0.015959097),
 ('ADM_RATE_ALL', 0.014765438),
 ('UGDS_WHITE', 0.013748974),
 ('COMP_ORIG_YR4_RT', 0.00748397),
 ('COMP_ORIG_YR2_RT', 0.0066271634),
 ('PPTUG_EF', 0.0061399858),
 ('INEXPFTE', 0.005556901),
 ('SATMTMID', 0.005271441),
 ('TUITIONFEE_PROG', 0.005024421),
 ('TUITIONFEE_IN', 0.0049120053),
 ('COMP_ORIG_YR3_RT', 0.0046591046),
 ('TUITFTE', 0.0040798658),
 ('HIGHDEG_Associate_degree', 0.0029162674),
 ('LOAN_COMP_ORIG_YR3_RT', 0.0025635613),
 ('UGDS', 0.0025409288),
 ('TUITIONFEE_OUT', 0.0024787188),
 ('UGDS_ASIAN', 0.0023418262),
 ('CONTROL_Private_nonprofit', 0.0021285913),
 ('COSTT4_P', 0.0021087895),
 ('AVGFACSAL', 0.0020022623),
 ('COUNT_WNE_P10', 0.0019879332),
 ('PCTFLOAN', 0.0017065941),
 ('UGDS_HISP', 0.0016695685),
 ('DEATH_YR4_RT', 0.0015539941),
 ('CONTROL_Public', 0.0015435874),
 ('UGDS_NHPI', 0.0015275691),
```



```
('UGDS_UNKN', 0.0014071261),  
( 'UGDS_NRA', 0.001291639),  
( 'UGDS_2MOR', 0.0011405724),  
( 'COSTT4_A', 0.0009246818),  
( 'ADM_RATE', 0.0008819474),  
( 'UGDS_AIAN', 0.00033204918)]
```

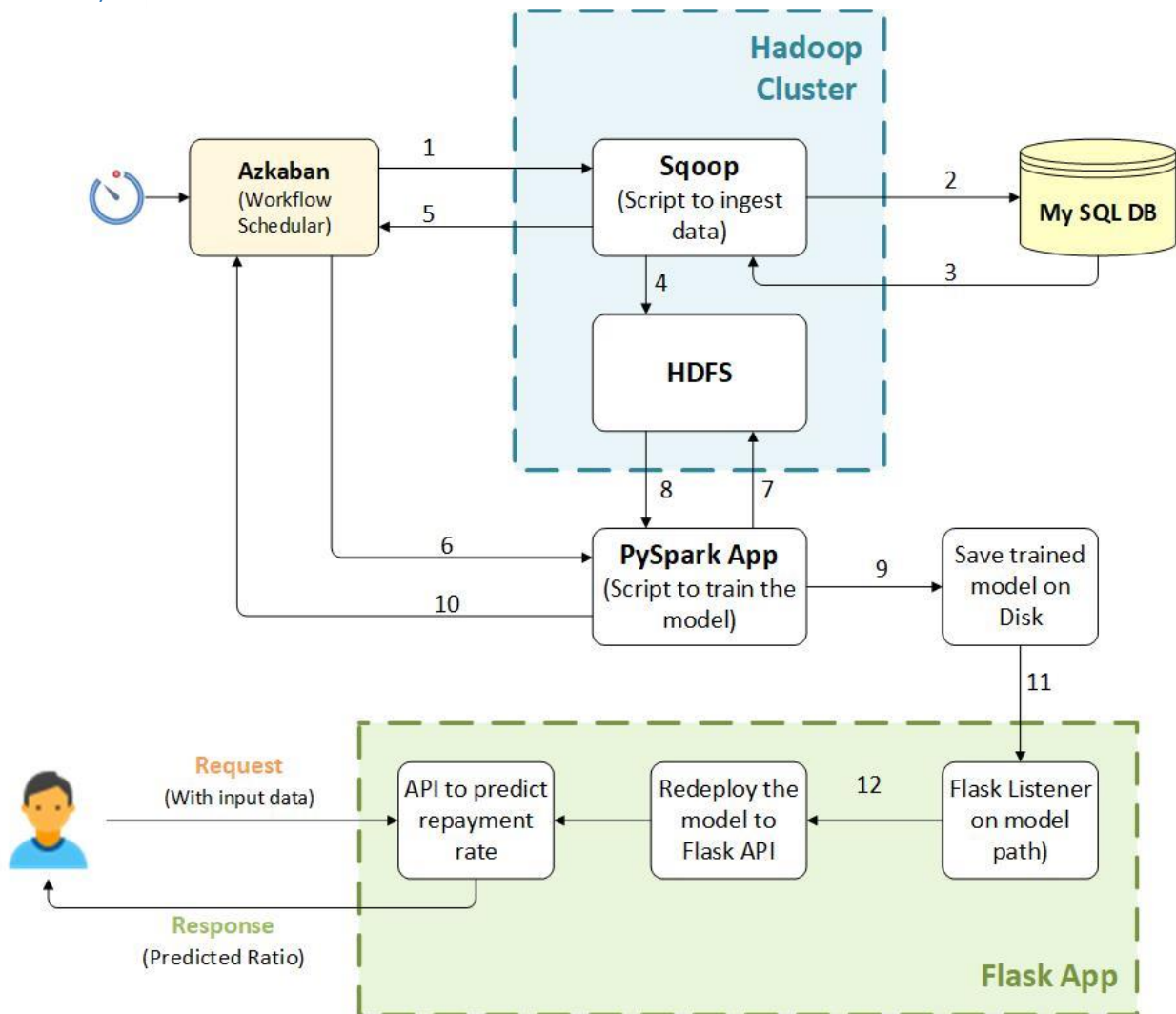
## 7. Deployment

### 8.1 System Main Components

- My SQL Database: Where raw Scorecard data for US institutions was stored.
- Azkaban: Is a batch workflow job scheduler created at LinkedIn to run Hadoop jobs. It resolves the ordering through job dependencies.
- Sqoop: Apache application used for transferring data between relational databases and Hadoop HDFS for MapReduce processing and so on.
- HDFS
- Spark App.
- Flask: A web framework provides some tools, libraries and technologies that allow to build a web application in python.  
Flask application is a portal where end user can submit his request and show the response



## 8.2 System Workflow



- 1: Azkaban scheduler triggers Sqoop script every 15 min
- 2: Sqoop script starts to ingest chunk of data from My SQL RDMS
- 3: Return data chunk to Sqoop
- 4: Sqoop save returned data to HDFS
- 5: Sqoop acknowledge Azkaban that its task is completed
- 6: Then Azkaban scheduler triggers PySpark App to start data processing task
- 7 , 8: PySpark App executers drill data from HDFS , and execute training jobs using the best parameters for XGBoost model
- 9: PySpark App executers save the trained model to the disk
- 10: PySpark App acknowledge Azkaban that its task is completed
- 11 , 12: Flask listener listen to the trained model updates saved on the disk , Then redeploy the updated model to Flask API

## 8. End User App

The interface of the end user app is as below.

Miners Model Client Interface

Enter Model Hyperparameters

AVERAGE AGE OF ENTRY AGE_ENTRY	AVERAGE SAT SAT_AVG_ALL
MATH SAT SCORES SATMTMID	NUMBER OF UNDERGRADUATES UGDS
NUMBER OF NON-WORKING STUDENTS COUNT_NWNE_P10	NUMBER OF WORKING STUDENTS COUNT_WNE_P10
MEAN EARNINGS OF WORKING STUDENTS MN_EARN_WNE_P10	MEDIAN EARNINGS OF WORKING STUDENTS MD_EARN_WNE_P10
NET-TUITION REVENUE PER FTE STUDENT TUITFTE	INSTRUCTIONAL EXPENDITURES PER FTE STUDENT INEXPFTE
IN-STATE TUITION AND FEES TUITIONFEE_IN	OUT-OF-STATE TUITION AND FEES TUITIONFEE_OUT
PROGRAM-YEAR TUITION AND FEES TUITIONFEE_PROG	AVG COST OF ACADEMIC-YEAR ATTENDANCE COSTT4_A

The user fills the information he has about the institute in question and press on Build Models.

UGDS_ASIAN	UGDS_AIAN
NUMBER OF NATIVE HAWAIIAN UNDERGRADUATES UGDS_NHPI	NUMBER OF UNDERGRADUATES WITH 2/MORE RACES UGDS_2MOR
NUMBER OF NON-RESIDENT UNDERGRADUATES UGDS_NRA	NUMBER OF UNDERGRADUATES WITH UNKNOWN RACE UGDS_UNKN
% UNDERGRADUATES RECEIVING FEDERAL LOANS PCTFLOAN	% UNDERGRADUATES RECEIVING PELL GRANT PCTPELL
% PART-TIME UNDERGRADUATES PPTUG_EF	COMPLETION RATE WITHIN 2-YEARS COMP_ORIG_YR2_RT
COMPLETION RATE WITHIN 3-YEARS COMP_ORIG_YR3_RT	COMPLETION RATE WITHIN 4-YEARS COMP_ORIG_YR4_RT
COMPLETION RATE WITHIN 3-YEARS WITH FEDERAL LOANS LOAN_COMP_ORIG_YR3_RT	PERCENT DIED WITHIN 4-YEARS DEATH_YR4_RT
CONTROL 2 - Private nonprofit	DEGREE TYPE 2 - Associate degree
Build Model	

Then he gets the probability of repayment for this institute (if a student took a loan and got enrolled in this institute, the probability of repayment is the result)

The screenshot shows a web browser window with the address bar displaying "127.0.0.1:5000/predict". The page title is "Miners Model Client Interface". Below the title, it says "prediction is: 0.5210390628492307". The main content area is a form titled "Enter Model Hyperparameters". The form is divided into two columns of input fields. The left column contains: "AVERAGE AGE OF ENTRY" (AGE\_ENTRY), "MATH SAT SCORES" (SATMTMID), "NUMBER OF NON-WORKING STUDENTS" (COUNT\_NWNE\_P10), "MEAN EARNINGS OF WORKING STUDENTS" (MN\_EARN\_WNE\_P10), "NET-TUITION REVENUE PER FTE STUDENT" (TUITFTE), "IN-STATE TUITION AND FEES" (TUITIONFEE\_IN), and "PROGRAM-YEAR TUITION AND FEES". The right column contains: "AVERAGE SAT" (SAT\_AVG\_ALL), "NUMBER OF UNDERGRADUATES" (UGDS), "NUMBER OF WORKING STUDENTS" (COUNT\_WNE\_P10), "MEDIAN EARNINGS OF WORKING STUDENTS" (MD\_EARN\_WNE\_P10), "INSTRUCTIONAL EXPENDITURES PER FTE STUDENT" (INEXPSTE), "OUT-OF-STATE TUITION AND FEES" (TUITIONFEE\_OUT), and "AVG COST OF ACADEMIC-YEAR ATTENDANCE".

## 9. References

- <https://collegescorecard.ed.gov/data/>
- <https://www.kaggle.com/kaggle/college-scorecard>
- <https://github.com/benhamner/us-college-scorecard>
- <https://www.kaggle.com/apollostar/which-college-is-best-for-you/report>
- <https://www.kaggle.com/apollostar/which-college-is-best-for-you-part2/report>
- [https://rpubs.com/random\\_Island/education-loan-repayment](https://rpubs.com/random_Island/education-loan-repayment)
- <https://www.americanprogress.org/issues/education-postsecondary/reports/2016/12/19/295187/sharing-the-risk/>