Kaleb Cervantes
Vidya Pingali
Felicia Kuan

College Machine Learning Project Documentation

Table of Contents

Introduction

Financial aid is funding exclusively for college students in US. Financial aid is available from federal, state, educational institutions, and private agencies, and can be awarded in the forms of grants, education loans, work-study and scholarships. In order to apply for federal financial aid, students must first complete the Free Application for Federal Student Aid (FAFSA). Colleges then use a student's FAFSA along with the college's resources to determine the aid given to a student.

We think this is interesting as college admissions is very relevant, especially in the wake of all the college scandals going on. Additionally, student loans and student debt is a very hot topic and we want to get a deeper insight into how colleges assign grants to students. Many students end up going to a particular college solely based on the aid and scholarships it gives them, regardless of whether or not they wanted to go there. Many students across the nation depend on financial aid to go to college.

Santa Clara University is a very expensive school, costing approximately $70,000 a year. Since it is a private university, we know many students receive some sort of financial aid to attend here. It would be interesting to us, as SCU students, to know how our school gives financial aid to its students and see how it compares to other schools. It would be interesting to compare SCU's financial aid process to other schools, such as UCs, other private schools similar in size, or other Jesuit schools.

We want to predict how much percentage of their accepted freshmen students a college will give financial aid to based on a number of factors such as location, tuition feeds, religious affiliation, size, etc. We used the IPEDS dataset found on Kaggle/Tableau that gives us information about all universities in the U.S.

Before we began the project, we wanted to see what sort of conclusions we could draw from the data. Most importantly, we wanted to know whether certain attributes of a college/university influenced the percent of the student body who eventually receive any form of financial aid. Below in Figure 1.1, we observed that there was a significant difference between the averages public and private universities when it comes to the percentage of students receiving financial aid. That means the Control of Institution correlates in some way, making Control of Institution one of the attributes we care about. However, the question we found ourselves immediately addressing: how many attributes matter to us?

Data Preprocessing

They say the most important step is cleaning our data, so multiple parts of our projects were dedicated to deleting and renaming things. Just as a count, we had 144 attributes in the beginning, and this is ALOT to work with, so we wanted to remove a couple (or more).

Deleting Based on Correlation

We wrote a function that automates our preprocessing procedure by eliminating attributes whose correlation is too high to another attribute. The function called **droplist** returns a list of attributes too closely related to each other for us to drop.

But first, we need to define what is dropped. If two attributes are correlated too similarly (defined by their correlation exceeding our threshold), **droplist** will keep the first attribute and drop the second. This is because, if two attributes have too high of a correlation, this means that they will both inform us the same thing about the data, and it is a wasteful computation to keep two attributes that are essentially identical -- and slow our computation -- when only one of them will suffice.

The exception: we observed later that we had deleted an attribute called “Number of first-time undergraduates - total” and kept “Number of first-time undergraduates - in-state.” Although the attributes are essentially identical, for the purpose of our project we thought that the total one would be more useful for our project.

To still use **droplist** function, we added an additional dropping step (**ogdrop** -- which means original drop), where we preliminarily dropped attributes that will be identical to the one we want (so the “total”), so that the one attribute we want to keep will remain as an attribute we will observe, and the other stuff (the “in-state,” “foreign country,” etc) that will be identical to the total will just be dropped.

After we drop the attributes we specified in **ogdrop**, then we ran our correlation function. In total, from this function, we dropped **44 attributes** deemed too similar to each other to matter.

Hand-Picking Attributes to Drop

Then, we went through the list of remaining attributes and used common sense to eliminate the attributes that are completely unrelated to our project about allocation of financial aid to undergraduates-- for example longitude and latitude.

The attributes we picked to drop are saved to **nextdrop**, and in the next section we will go through and explain (in sections) why these attributes are useless to our analysis.

A. Location-based Attributes

These are attributes linked to the geographical location of the school, ie. longitude. The geography or level of urbanization is information about the environment around the school has no relation to how much financial aid a student is allocated, nor does it relate to how much resources the university has to allocate. Thus, these attributes are useless to our analysis.

- ➢"Longitude location of institution"
- ➢"Latitude location of institution"
- ➢"Geographic region"
- ➢"Sector of institution"

➢"Level of institution" ➢"Tribal college"
➢"Degree of urbanization (Urban-centric locale)" ➢"FIPS state code"
➢"Endowment assets (year end) per FTE enrollment (GASB)"
➢"Endowment assets (year end) per FTE enrollment (FASB)"

B. Duplicate Information

Some attributes we observed by hand will inform us the same thing about a university as another attribute. Most of these attributes should have (theoretically) been removed by our drop function because the correlation would be too high. Alas, in practice, we will still have to clean things up. For example, we removed the attribute “County name” because it’s linked to the “ZIP code” that we kept.

C. Graduate or Beyond

Our analysis is on undergraduate students attending 4-year colleges, so we decided to remove all attributes relating to graduate students. In addition, we also didn’t care about the scrubs attending a junior college, so we removed those too (perhaps this deserves a section of its own).

What’s convenient is that some of these attributes are misspelled and parsed weird (using two spaces instead of one), so perhaps it’s a good thing that we’re deleting them.

➢"Doctor's degrees - professional practice awarded” ➢"Offers Associate's degree"
➢"Offers Master's degree" ➢"Associate's degrees awarded"
➢"Offers Postbaccalaureate certificate" ➢"Offers Post-master's certificate"
➢"Offers Doctor's degree - research/scholarship" ➢"Offers Doctor's degree - other"
➢"Offers Doctor's degree - professional practice" ➢"Offers Other degree"
➢"Doctor's degrees - other awarded" ➢"Highest degree offered"
➢"Postbaccalaureate certificates awarded" ➢"Post-master's certificates awarded"
➢"Doctor's degrese - research/scholarship awarded"
➢"Number of students receiving a Postbaccalaureate or Post-master's certificate"

D. Student Enrollment Information

This is information useful for enrolling students, particularly if we were interested in the profiles of students who were accepted/rejected/decided not to come. However, we’re only interested in students who ended up attending the university because these are the students who wind up using the university’s financial aid. Therefore, we only care about information enrolled students, not admitted and others.
In addition, we don’t care about the student’s test scores because this mostly applies to college applications, not financial aid.

➢"Applicants total" ➢"Percent of freshmen submitting SAT scores"
➢"Admissions total" ➢"Percent of freshmen submitting ACT scores"
➢"Admissions yield - total"

E. Part-Time Students

We also ignored part-time students because we only focus on full-time students. Therefore, we will also ignore students who receive certificates for completing less than four years of university because

these are also not full-time students. Universities are much less likely to allocate financial aid to students who are not full-time.

➢"Certificates of less than 1-year awarded"
➢"Estimated enrollment, part time"
➢"Offers Less than one year certificate"
➢"Estimated freshman enrollment, part time"
➢"Offers Two but less than 4 years certificate"
➢"Estimated graduate enrollment, total"
➢"Certificates of 1 but less than 2-years awarded"
➢"Certificates of 2 but less than 4-years awarded"
➢"Offers One but less than two years certificate"
➢"Number of students receiving a certificate of 1 but less than 4-years"
➢"Estimated graduate enrollment, part time"

F. Financial Aid Information

You might be thinking, "Felicia, are you crazy? Why would you remove attributes relating to financial aid in a project analyzing financial aid?!"

And to that, we contend that according to our correlation function, all of these attributes relating to the percent of freshmen receiving any sort of financial aid are defined as similar enough because their correlation exceeded the correlation threshold function we defined previously. Therefore, it's only necessary to keep 1/7 of these.

➢"Percent of freshmen receiving federal grant aid"
➢"Percent of freshmen receiving other federal grant aid"
➢"Percent of freshmen receiving state/local grant aid"
➢"Percent of freshmen receiving institutional grant aid"
➢"Percent of freshmen receiving student loan aid"
➢"Percent of freshmen receiving other loan aid"

G. Unknown Metrics

These were attributes that served as ranks within the data, but they first, don't really relate with our analysis about financial aid, and second of all, we wouldn't know how a higher ranking on an arbitrary (to us) metric for measuring the quality of a school relates to financial aid. So, we just removed them.

➢"Carnegie Classification 2010: Basic"
➢"Level of institution"

H. Graduate Student Identity

Similar to section C, this section deletes the attributes relating graduate student information, but this time about their profile and identity. As in Section C, because this is information regarding graduate students, it will be removed.

➢"Percent of graduate enrollment that are American Indian or Alaska Native"
➢"Percent of graduate enrollment that are Asian"
➢"Percent of graduate enrollment that are Native Hawaiian or Other Pacific Islander"
➢"Percent of graduate enrollment that are White"
➢"Percent of graduate enrollment that are two or more races"
➢"Percent of graduate enrollment that are Race/ethnicity unknown"

➢"Percent of graduate enrollment that are Nonresident Alien"
➢"Percent of graduate enrollment that are women"

Result: 58 attributes dropped.

Thus, the 27 attributes we have left are:
ID number, Name, ZIP code, Religious affiliation, Offers Bachelor's degree, Enrolled total
SAT Critical Reading 25th percentile score, Percent admitted - total, Tuition and fees, 2010-11, State abbreviation
Control of institution, Historically Black College or University
Percent of undergraduate enrollment that are American Indian or Alaska Native
Percent of undergraduate enrollment that are Asian
Percent of undergraduate enrollment that are Black or African American
Percent of undergraduate enrollment that are Hispanic/Latino
Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander
Percent of undergraduate enrollment that are White
Percent of undergraduate enrollment that are two or more races
Percent of undergraduate enrollment that are Race/ethnicity unknown
Percent of undergraduate enrollment that are Nonresident Alien
Percent of undergraduate enrollment that are women
Percent of first-time undergraduates - in-state
Percent of first-time undergraduates - foreign countries
Percent of first-time undergraduates - residence unknown
Graduation rate - Bachelor degree within 4 years, total
Percent of freshmen receiving any financial aid or degree within 4 years, total

Removing Rows

We're looking at the "Offers Bachelor's Degree" and if the college for that attribute has the value "Implied No", we know that the university only offers graduate degrees or higher. We don't need these colleges because they're related to graduate schools, and we are only analyzing the undergraduate universities. Thus, we removed these universities that only offer graduate degrees or above.

Then, we noticed that of the remaining attributes, all values in this "Offers Bachelor's Degree" attribute were "Yes," and since this column has uniform values, it's unnecessary to keep this column.Thus, at this point, we have **removed 13 rows**, with **1492 colleges remaining**, and **26 attributes remaining**.

Dealing with Null Values

We noticed a problem, that for some of the attributes, an unusually high percentage of the colleges have null values for these attributes we'll be observing. This is problematic, as too many of these null values in an attribute will skew our analysis. To judge how many null values was considered "too many," we took a count of all 1522 rows and observed how many null values there were for each of the 26 attributes and calculated a percentage. The following is the count:

ID number	1522 non-null int64
Name	1522 non-null object
ZIP code	1522 non-null object
Religious affiliation	1522 non-null object
Enrolled total	1377 non-null float64
SAT Critical Reading 25th percentile score	1169 non-null float64
Percent admitted - total	1376 non-null float64

Tuition and fees, 2010-11	1490 non-null float64
State abbreviation	1522 non-null object
Control of institution	1522 non-null object
Historically Black College or University	1522 non-null object
Percent of undergraduate enrollment that are American Indian or Alaska Native	1522 non-null float64
Percent of undergraduate enrollment that are Asian	1522 non-null float64
Percent of undergraduate enrollment that are Black or African American	1522 non-null float64
Percent of undergraduate enrollment that are Hispanic/Latino	1522 non-null float64
Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander	1522 non-null float64
Percent of undergraduate enrollment that are White	1522 non-null float64
Percent of undergraduate enrollment that are two or more races	1522 non-null float64
Percent of undergraduate enrollment that are Race/ethnicity unknown	1522 non-null float64
Percent of undergraduate enrollment that are Nonresident Alien	1522 non-null float64
Percent of undergraduate enrollment that are women	1522 non-null float64
Percent of first-time undergraduates - in-state	911 non-null float64
Percent of first-time undergraduates - foreign countries	911 non-null float64
Percent of first-time undergraduates - residence unknown	911 non-null float64
Graduation rate - Bachelor degree within 4 years, total	1476 non-null float64
Percent of freshmen receiving any financial aid	1492 non-null float64

A. 40% Null

The most alarming revelation when doing this summation was seeing how the following attributes

➢"Percent of first-time undergraduates in-state"

➢"Percent of first-time undergraduates foreign countries"

➢"Percent of first-time undergraduates residence unknown"

These three had 40% of the colleges record null values for these attributes, so we decided to drop these three because 40% of missing values will most definitely skew our results. This leaves us with **26 attributes**.

B. Percent of freshmen receiving any financial aid

This is the most important attribute in our dataset because this is our result we are observing and trying to analyze which attribute affect the value of this attribute, it wouldn't be very useful to us in this analysis if the value in this attribute were null. Thus, we removed the **30 rows** that were null for this attribute; the amount of rows we're analyzing decreased from 1522 to **1492 rows**.

C. SAT Critical Reading Score

We also noticed that a little over 20% of the values for the attribute

➢"SAT Critical Reading 25th percentile score"

were null. It is quite logical that this is so, as some schools in the U.S. accept ACT scores instead of SAT. Although we were concerned that by removing this attribute, we would not be observing how merit and the academic success (defined only by the SAT score) influence the allocation of financial aid. Arguably, this is an important attribute to keep, but with 20% null values we wouldn't be predicting any results accurately.

Perhaps, if later in the project we decide that this attribute is interesting, we can recover the data attached to this attribute and remove the null rows and re-do our analysis. However, then this would narrow our estimation to only colleges that accept SAT scores.

Thus, we remove one more attribute, leaving us with **24 attributes**.

D. 8% Null: Enrolled Total and Percent Admitted

Enrolled Total has 8% null values in our dataset and we realized that by looking at the original, undeleted set of data, Enrolled Total has the same values as Total Enrollment, yet Total Enrollment has ZERO null values. Thus, we decided to go back before we deleted all the attributes from the correlation functions and dropped Enrolled Total so that Total Enrollment would be kept during the correlation step.

E. List of Attributes used for Analysis

Finally, we've arrived at the point where we've selected which attributes we will use in our analysis.

We are confident that as for the remaining null values, they apply to below 5% of the colleges (rows/data), so their effect on the analysis should be inconsequential. Thus, the final attributes we kept using this procedural data pre-processing and their count of non-null objects are as follows:

ID number	1492 non-null int64
Name	1492 non-null object
ZIP code	1492 non-null object
Religious affiliation	1492 non-null object
Enrolled total	1373 non-null float64
SAT Critical Reading 25th percentile score	1167 non-null float64
Percent admitted - total	1373 non-null float64
Tuition and fees, 2010-11	1488 non-null float64
State abbreviation	1492 non-null object
Control of institution	1492 non-null object
Historically Black College or University	1492 non-null object
Percent of undergraduate enrollment that are American Indian or Alaska Native	1492 non-null float64
Percent of undergraduate enrollment that are Asian	1492 non-null float64
Percent of undergraduate enrollment that are Black or African American	1492 non-null float64
Percent of undergraduate enrollment that are Hispanic/Latino	1492 non-null float64
Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander	1492 non-null float64
Percent of undergraduate enrollment that are White	1492 non-null float64
Percent of undergraduate enrollment that are two or more races	1492 non-null float64
Percent of undergraduate enrollment that are Race/ethnicity unknown	1492 non-null float64
Percent of undergraduate enrollment that are Nonresident Alien	1492 non-null float64
Percent of undergraduate enrollment that are women	1492 non-null float64
Graduation rate - Bachelor degree within 4 years, total	1475 non-null float64
Percent of freshmen receiving any financial aid	1492 non-null float64

Encoding Data

What needs to be encoded? Before we fit the data, we need to encode our attribute values that are categorical and change them into numerical. So far, it is the following attributes that need to be numericized.

- ➢ZIP code
- ➢Religious Affiliation
- ➢State abbreviation
- ➢Tuition and fees, 2010-11
- ➢Control of institution
- ➢Historically Black College or University

While we're encoding and began to try fitting data, we noticed several problems with the way some of these attributes are encoded.

Unique Values

A. Name and ID

When we're creating a model to predict values, it's not useful to have all unique values for an attribute because there would be nothing to predict. Thus for our training and testing purposes, we removed name and ID number when we're training because they're unique values.

B. Religious Affiliation

We took a closer look at the values for Religious Affiliation and realized that the values were far too specific-- we had the typical Catholic and Protestant universities, but then we had like 3 Mormon, and 10 other small denominations of Christianity that had one or two universities each, and this made the classification of our data too specific and predictions less accurate. Therefore, we went through and wrote a function to encode this attribute separately. If the school is religiously affiliated to any religion, it received a value of 1, and if not, 0.

C. ZIP code

Similarly to name and ID, many of the colleges don't share a **ZIP code** with another university, so ZIP code becomes a unique value. It skewed the accuracy of our prediction because we weren't able to classify all 1400 unique ZIP codes. However, we didn't want to just remove them because we've removed a lot of attributes already, so we decided that other attributes we've deleted before might be able to replace it with better results.

We considered the attribute **County name** as a replacement for ZIP code, but it wouldn't tell us particularly useful, as some county names are used multiple times across different states. We concluded that the only attribute relating to a region a university is located in would be the attribute called **"Degree of Urbanization."**

D. Adding Urbanization

When we went back to our code to add the **Urbanization** attribute, we noticed that Urbanization had 12 different values:

City: Large	294	Suburb: Large	290
City: Small	225	Suburb: Midsize	49
City: Midsize	185	Suburb: Small	36
Town: Distant	165	Rural: Fringe	58
Town: Remote	125	Rural: Distant	30
Town: Fringe	62	Rural: Remote	15

Particularly for the categories with less than 100 values, it wouldn't be useful to have the categories split so specifically, as it would cause overfitting. Thus, we wrote code to merge a few of the categories. Our new specification merged all the Suburbs in one, and all the Rural in one category. Then, we made Fringe and Remote Towns in one.

City: Large	294	Suburb:	375
City: Small	225	Rural:	103
City: Midsize	185	Town: R+F:	187
Town: Distant	165		

After altering the way urbanization is encoded, the training and testing data became more accurate. We tested this and by altering the way the encoding is done, training data became more accurate.

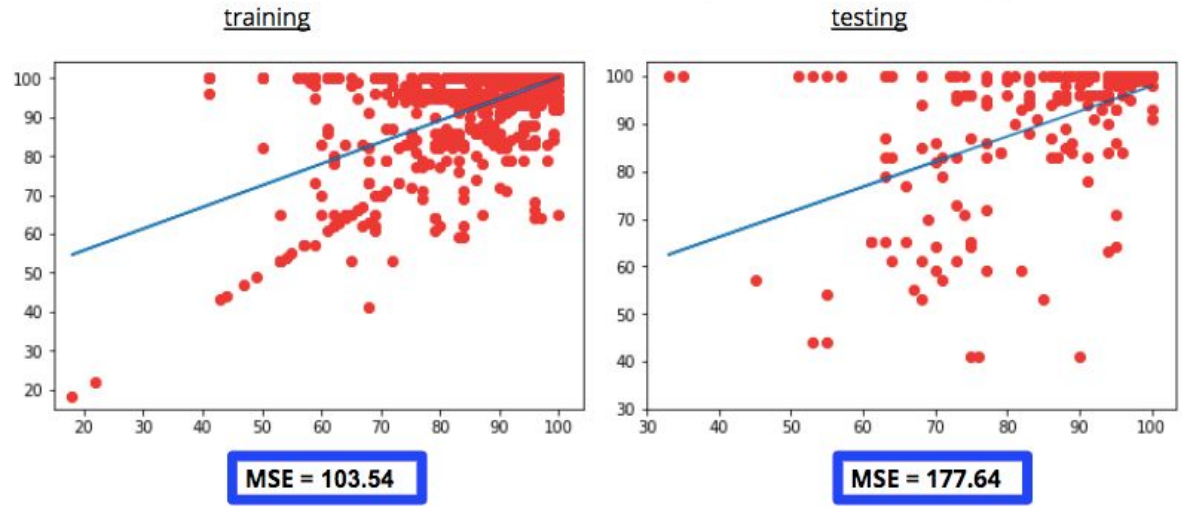
Experimenting

Now we're ready to fit the data! First, what we know about our data that it's quantitative. Therefore, we can only use **regression** algorithms and not classifier ones (like KNN) because classifiers deal with qualitative data, not ours.

Note: for ALL graphs, the x axis = actual value
the y axis = predicted value

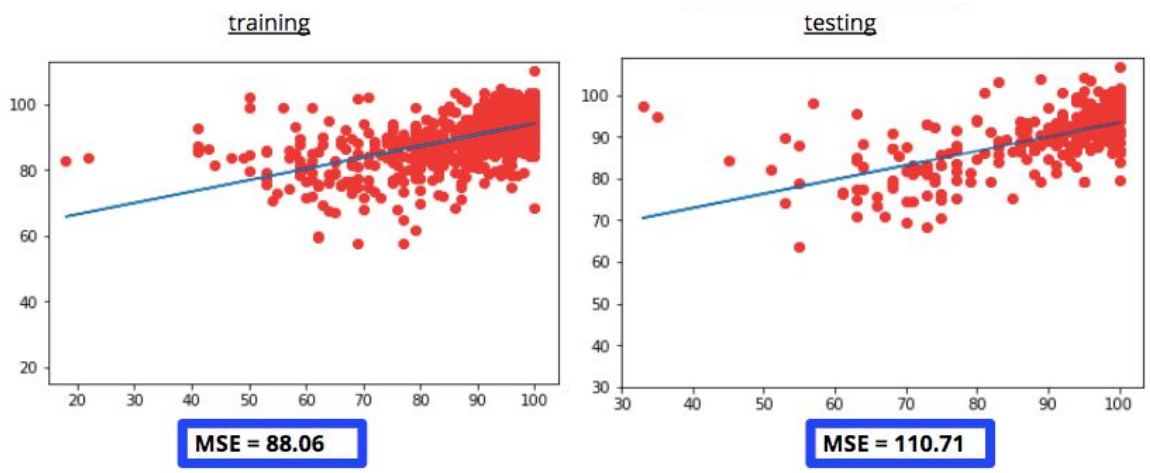
Logistic Regression

We attempted to fit our data using a logistic regression model, but as you can see for our training and testing data below, the points do not lie on the line at all. Interestingly, for the training results, we can see that at one point, there are points that line up in a diagonal line slightly below the ideal line, but testing data shows that the line has no effect on the accuracy of the regression. By far, logistic regression was our worst method yet.



Linear Regression

We also attempted to fit out data using linear regression. While it gave less of an error than logistic regression, it's still clearly not the best model and KNR is superior.

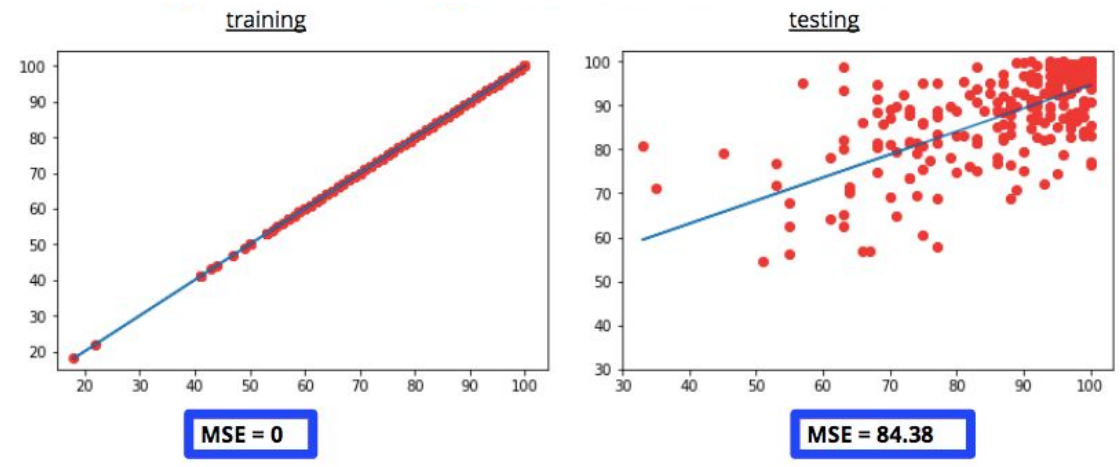


KNN Regression with Weighted Distance:

What is KNN Regression? This is basically a regression algorithm that uses KNN to tweak the regression a little.

With this in mind, we attempted to use KNN regression with a closest distance priority. Normally, KNN takes its k nearest neighbors and weighs them the same regardless of the distance, but for this KNN with a closest neighbors priority, this means that when choosing nearest neighbors, the algorithm would prioritize choosing the k nearest neighbors and weighs the closer ones heavier than the ones farther away. Thus, when a neighbor whose distance is 0 away, the algorithm basically only cares about this point, not the others, which makes KNN useless because in this case it only looks like the one or two other neighbors who are literally the same thing.

This is a major problem because clearly, this algorithm will severely overfit (take a look at the Training data below). And this algorithm uses k=5 neighbors.



Reasons it overfit:

1. Since the algorithm weighting points closer to it higher in the training set, so because of this there's a neighbor on it where the distance between it and the neighbor is closer to 0. So we'll always pick that closest neighbor. We need to make a function that allows the training data to not pick the point whose distance is to 0.
2. **Regression based on distance would always overfit** → would be nice if we had an algorithm that would train a model and not choose the neighbor whose distance is 0 (would be nice to spend more time writing this algorithm!)
3. To find the nearest distance we noticed that Manhattan distance gave us the lowest test error, but honestly the distance one gave us so many errors, the type of distance didn't really make a difference, so we still used Euclidean.

Current Observations and Speculations

Schools that give less financial aid (less than 60% of the students receive any form of financial aid) are overestimated by A LOT even more in our model, and the prediction error is really high for these school. Why is this happening?

We don't have enough data points for schools that give less financial aid (give less than 60% of financial aid) in the dataset, as we only have about 50 of these schools. Therefore, when we use KNN, the prediction will take its k nearest neighbors, but its neighbors are more likely to be the schools that provide more financial aid, so when predicting the schools with lower, our estimation will gravitate towards the higher prediction. Thus, our model will always **overestimate** (by a lot) the schools with less financial aid. You take a look at the testing data above to see how inaccurate the prediction is for the lower financial aid.

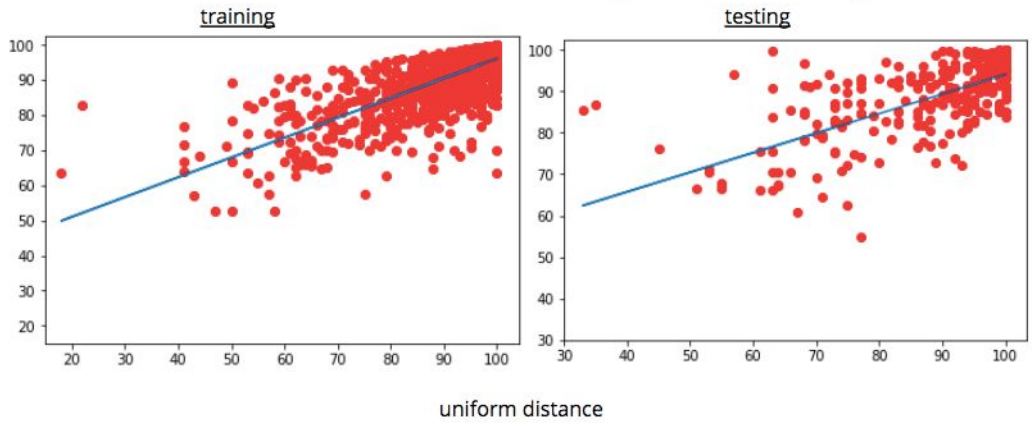
In terms of the universities that give much more financial aid, our models predict these much better, which tell us that we really just need more data for the schools that give 60% or less of the students financial aid. We have the most accurate prediction for schools that give more financial aid (80% or more).

Therefore, if we had more time, it would be interesting to modify the algorithm for weighing nearest neighbors to **not** choose the neighbor whose distance from the point is zero. Doing this will take much more time because we'd have to look further into the documentation for ScikitLearn and make sure what we're doing is correct.

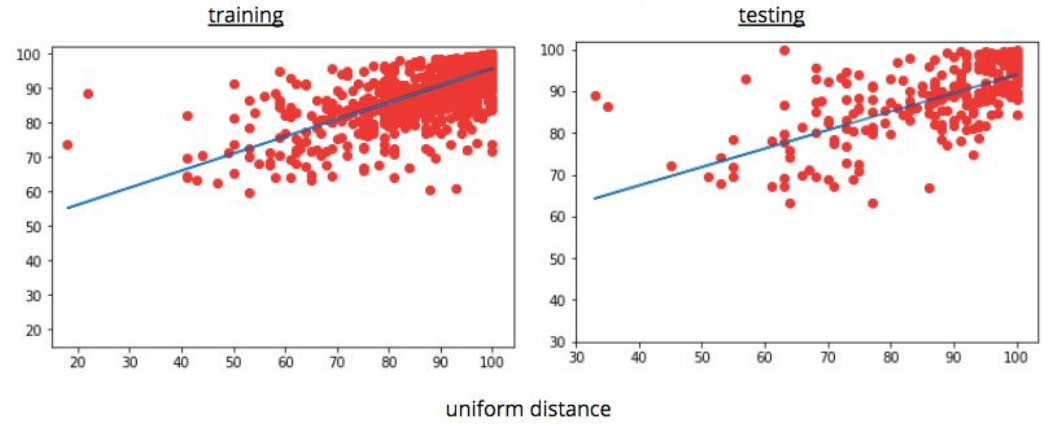
KNN Regression with Uniform Distance

In other words, this is just simple KNN Regression because we realize that the distance one will always overfit. In this section, we altered the number of neighbors to see whether or not it reduces the error. Ideally it should look more like a straight line

KNR with k = 5

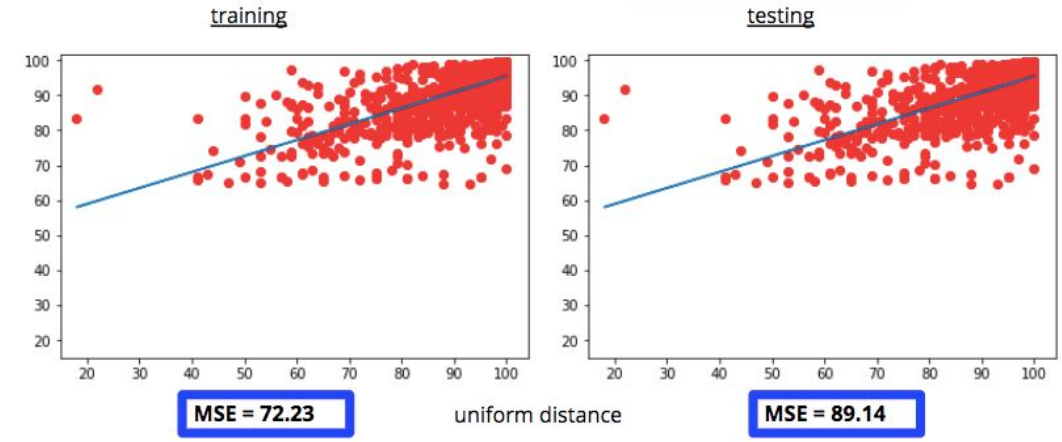


KNR with k = 10

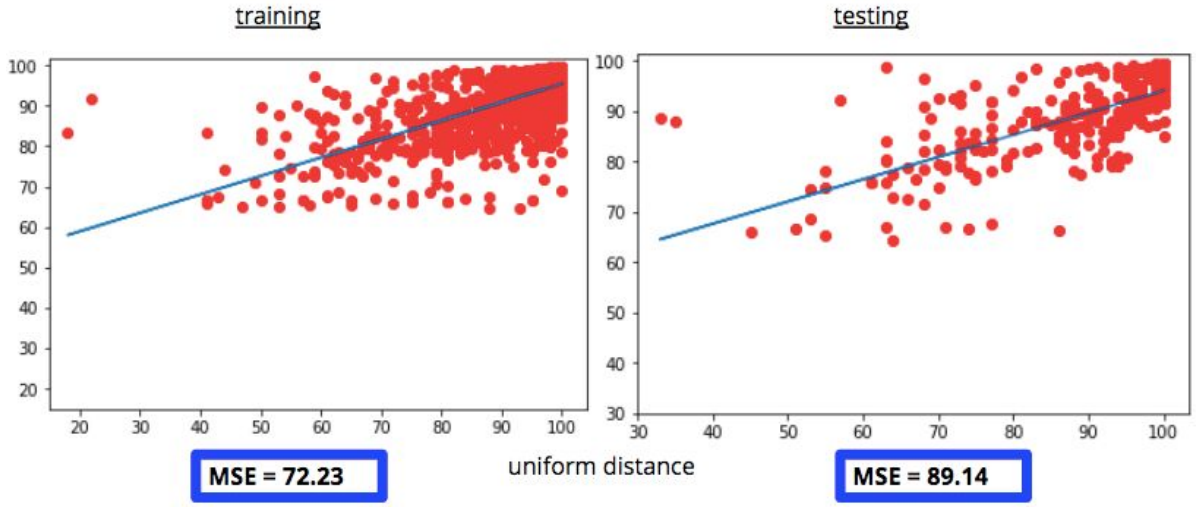


More neighbors increase accuracy (corr + and MSE -) but it slows down our computation time

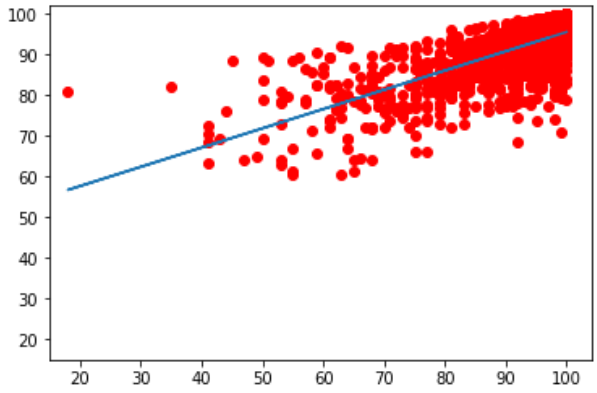
KNR with k = 20



KNR with Urbanization Encoding; k = 20



KNR with k = 20 gave the least error so that's the best model.

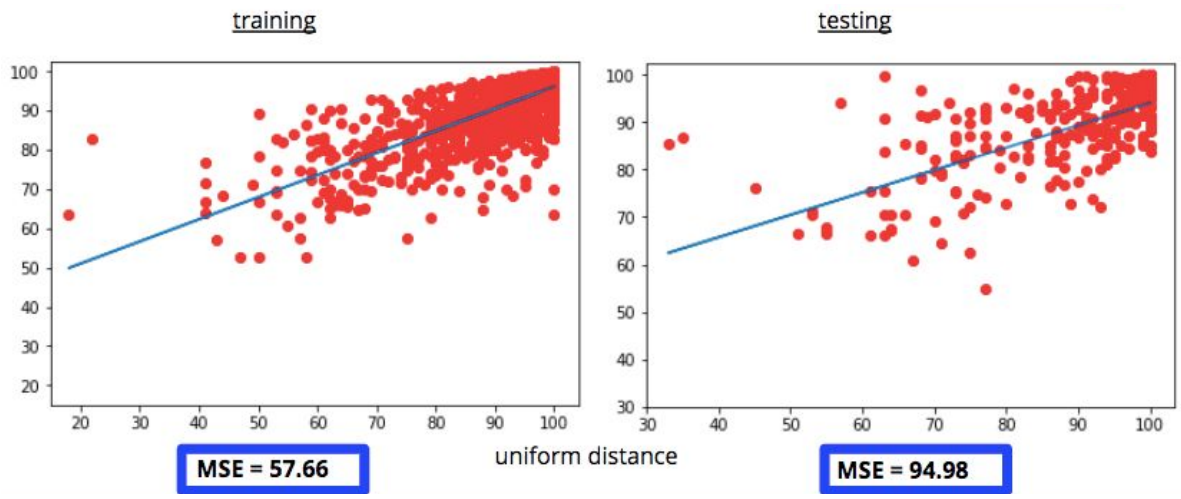


Training data with default settings with sklearns encoding
MSE = 110.71474396781576

Removing Race-related Attributes

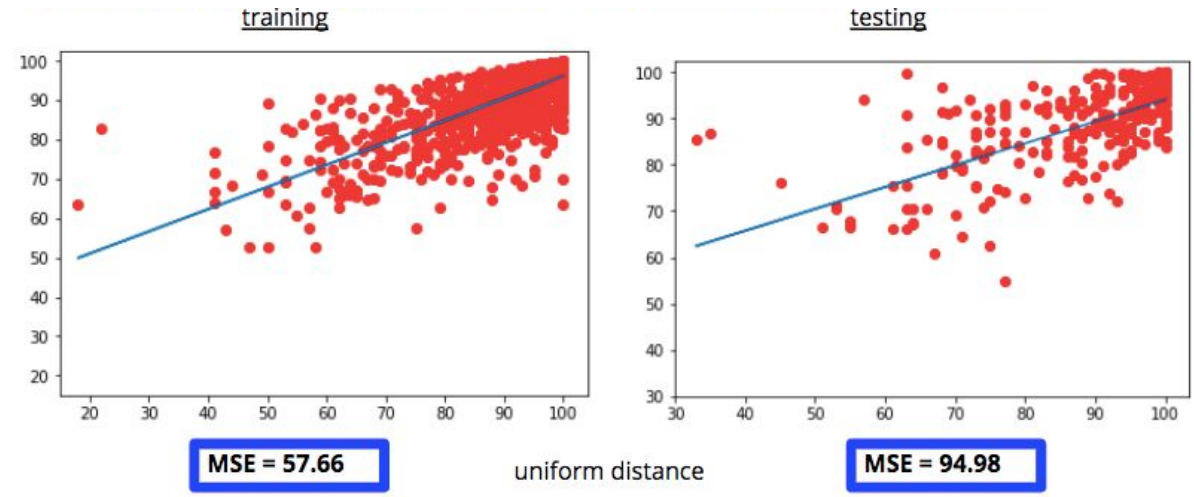
We went back to our original attribute list that we finalized from KNR and dropped the 8 attributes that corresponded to race to see how race affects financial aid statistics (refer to the graphs below). Interestingly, removing all attributes relating to race decreases the error for our training data, but increased the error in the testing data.

This shows that the race attributes are very important in determining financial aid, because removing these will reduce the accuracy of our algorithm. This makes sense, as race can be correlated to factors such as socioeconomic status, it is very relevant to financial aid statistics.



Removing Religious Affiliation Attribute

We went back to our original attribute list and dropped the religious affiliation attribute to see how it has an effect on our prediction. Since SCU is a Jesuit school, we thought it would be interesting to analyze this. It was very similar in results to removing the race attributes as you can see by their graphs. This indicates that religious affiliation of a university is also very important in determining the amount of financial aid it is capable of giving, as removing this attribute increased the error of our prediction.



Conclusion

In conclusion, our model didn't fit our data that well. As mentioned before, we didn't have a lot of data for the schools that give financial aid to less than 60% of their students so it was hard for these points to find accurate neighbors, as they constantly gravitated towards the neighbors with higher values.

Further Research

A. Better Data Pre-Processing?

Given more time, it would be useful to study how the attributes we chose truly affect financial aid, as we are not completely sure how these attributes correlate with the result. Did we eliminate an attribute that actually has an effect on financial aid? Did we keep an attribute that has little correlation? Did we eliminate too many things when trying to reduce our null values? Only further work on this project can tell. We only could do the best we can to pre-process the data in a fair and reasonable manner, but perhaps we made mistakes here and there. Given more time to research this subject, we will have to back up our reasons for eliminating an attribute with more concrete evidence-- perhaps something graphical.

B. Expanding on KNN Regression

If we had more time, it would be interesting to modify the KNN Regression based on Weighted Distance algorithm for weighing nearest neighbors to **not** choose the neighbor whose distance from the point is zero. Because of our project's heavy reliance on multiple attributes that influence the end result, we believe that distance Doing this will take much more time because we'd have to look further into the documentation for Scikit Learn and make sure what we're doing is correct.

C. Exploring Other Algorithms

We would look into DBSCAN, which is density-based. Our data could benefit from density-based analysis, particularly because we have more data points for schools that give more financial aid than those that do not. We also researched a bit into Kernel Density but didn't understand how the algorithm worked so if we had more time we would look into that.

We would also look into elastic nest that uses iterative fitting. We can also run more iterations of our regression, as we found an algorithm that goes through several rounds, training the model with different data each time and achieves a better result.

D. 6 More Months

If we had 6 more months to improve and expand on our project, we would do the following:

- Find the optimal amount of neighbors (most accurate and least computation)
- Change the distance regression algorithm to choose the point that's not 0 distance
- Experiment with other regression algorithms that could be applicable
- Research more about a density-based regression
- Get more data on schools that give financial aid to 60% or less of its students