US Universities In-State Tuition Prediction

Data Description & Goal

The aim of this project is to utilise our data to predict the In-State tuition of American based universities.

The data we used for this project was collected by two independent organizations however they are related to the same universities hence may prove to be useful to the final we create. These datasets were as follows:

- *US News Universities data* – This dataset contains information on the various admission rates, SAT and ACT admission criteria, location, graduation rates, percentage of alumni who contribute as well as other relevant information on the universities.
- *AAUP Data* – This dataset contains information on the salary and compensation of the various universities to the assistant professors, full professors, and associate professors. Additionally, there is a host of other variables that provide more insight on faculty specific information. It is important to note that this data represents the rankings of the salaries and compensation and not the actual mean of the professor compensation packages.

As mentioned earlier, the data was collected by different sources/organizations hence in order to achieve what we considered our full dataset we had to merge these two sets of data. The merging of this data was based on the "FICE ID" variable. The FICE ID is a unique identification sequence that is assigned to a university by the Federal Government during incorporation.

The uniqueness of the FICE ID helped us identify the duplicate rows of data as will be discussed later on in our data exploration portion of this report.

The AAUP data contained 200 more rows of data than the US News data, this indicated that they collected data on more universities hence during merging some of the rows may not have corresponding FICE values and as such they would have to be discarded. Retaining only the rows of data that have corresponding values would also ensure that the missing values do not get out of hand and the granularity of the rows in the final dataset will be high.

The merged dataset had the following dimensions : 1134 rows with 49 variables. Of the 49 variables 3 of them were categorical and the remaining 46 were continuous variables.

Throughout this report we will walk you through our approach to predicting the In-State Tuition: this will entail the steps we took to better understand our data as well as our exploratory analysis process and variable selection methods for our models including model assessment.

Link to Data : http://lib.stat.cmu.edu/datasets/colleges

Data Preparation & Exploratory Analysis

The readme and metadata info attached to the dataset did not state what the collection methods were used to collect the data. In our quest to prepare the data for modelling the main issues we had to address were:

- Duplicate Rows
- Data Type conversions
- Missing Values
- Missing Values : KNN Imputation

Duplicate Rows

The datasets were merged based on the FICE values: this merging resulted in some of our data having duplicate rows or in other cases the data was conflicting for the various rows. We addressed such issues with the following approach:

Duplicate rows – Duplicate rows were easily corrected by removing the duplicate occurrences of the specific rows.

Duplicate Universities conflicting rows – These kinds of duplications were solved by removing the rows entirely as it would be hard to determine which of the various rows contained the accurate or most up to date information.

Erroneous FICE ID – An entry error concerning the Pennsylvania State University required us to search and update its FICE value since the ones provided by the two datasets were conflicting. It is worth noting that this was the only occurrence of such a case.

Datatype conversions

Upon loading the data, we realized all the data that we had were encoded as character types. However, this is not really the case as we could tell from the variable names and the description that aside from the three categorical variables in the data all the rest were meant to be numerical.

With the exception of university name, postal code of the university (state) and public/private status we converted the remaining variables to integers.

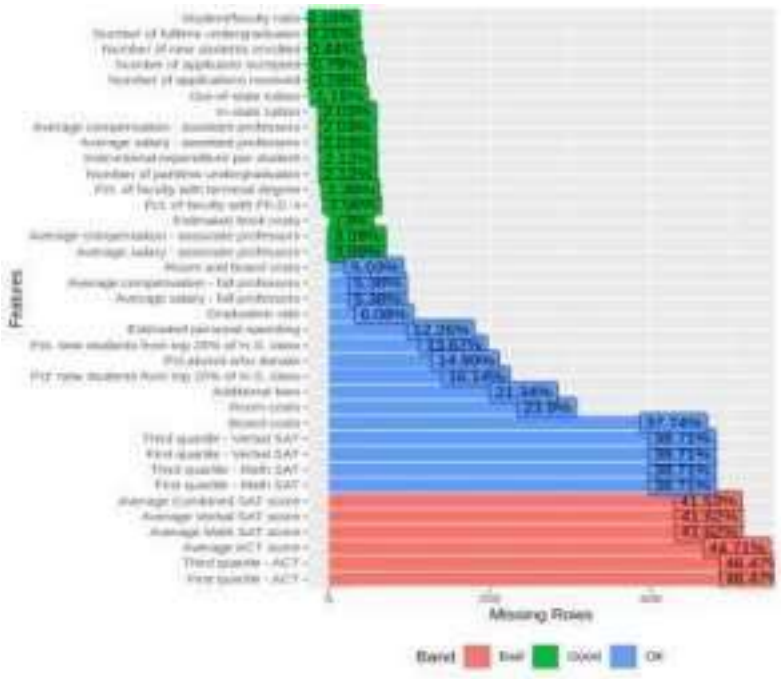
It is worth noting that the Public/Private status was converted to numerical as well but retained its status as a categorical variable with 0 representing public universities and 1 representing private universities.

Missing Values

This would be considered the main challenge with regard to the cleaning of this dataset. Some of the variables in our dataset had more than 40% of the values missing which would have heavily impacted the dataset. Any information provided on how the data was collected may help in improving or advising the methods we use to deal with the missing values.

Performing the datatype conversion before this step was essential to allow us to easily manipulate our data in a numerical form.

Below is a graph that shows the percentage of missing values across the variables in our dataset.



In our case we highlighted the missing value rate of our categories as follows:

Good – Variables missing less than 5% of the data.

OK – Variables that contain between 5% - 40% of their data.

Bad – Variables that are missing more than 40% of their data.

Our approach to missing values is defined based on the percentage of missing values:

Missing more than 40% of data - In such cases we decided that it would be in our best interests to remove these variables from our dataset, having them in the model would impact the accuracy and have a drastic effect on the coefficients of our model. Removing the missing rows with missing data meant that we would result in losing a decent amount of our data hence this was not considered as a valid approach to solving this issue.

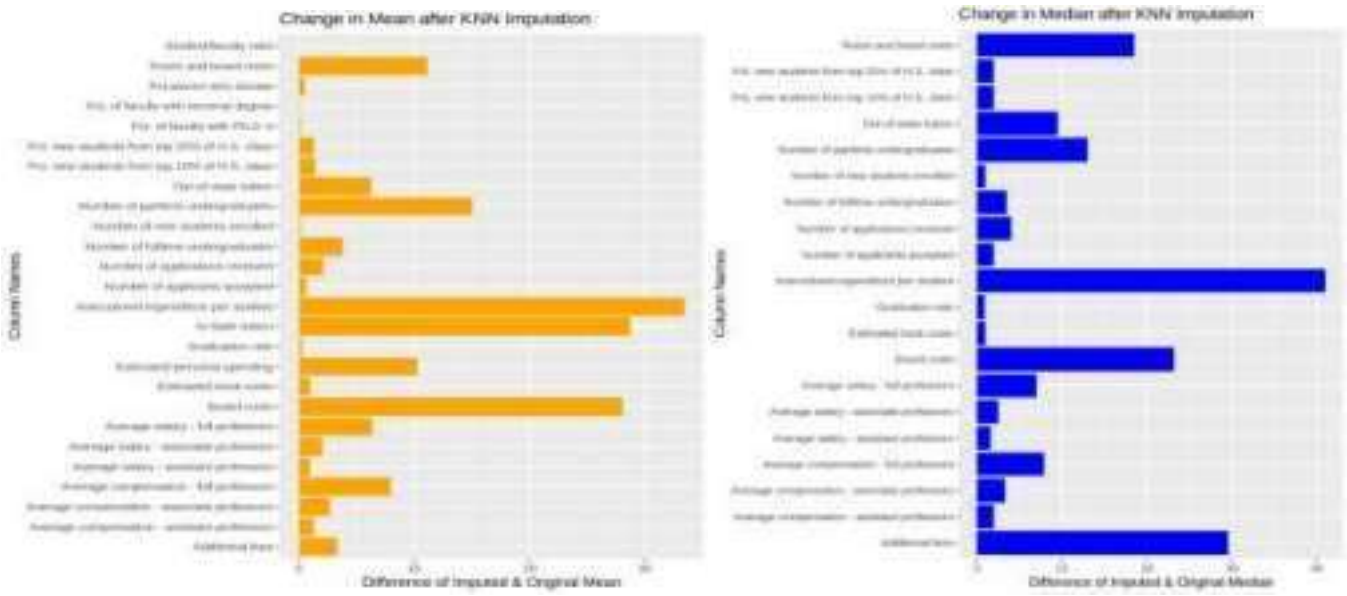
Missing less than 40% of data – Imputation was our chosen method to replace these missing values. The high range and variability in our features would make the replacing the missing values using the feature mean will cause sharp rises in the mean as the larger values will swing the mean to the right. Alternatively, the median of the feature may not always be the best representation of the data and as such we opted for a more elegant imputation system, *KNN Imputation.*

KNN Imputation

Following the reasons stated above we decided to use KNN imputation to replace our missing values. In our case any of the missing rows of data of a feature were replaced with the mean of

its ten nearest neighbours. Using this method ensured that the values we used for the replacement would result in a small difference when compared to the actual values were they available.

After imputation we analyzed the impact of the imputation on the absolute means and medians of the data. If the imputation worked correctly then the changes in mean and median will not be very drastic: the change being considered in this case is absolute hence it includes rises and drops in mean and median.



Changes in Median Imputation – Majority of the variables demonstrated only slight shifts in the median with the exception of the Instructional Expenditure per student and Additional Fees which registered the largest changes majorly due to the variability of the features and the lack of standardization in these charges hence the changes were large.

Changes in Mean Imputation – Just as we noted in the median category the changes for this segment as well were minimal for majority of the features in this case with the exception of Instructional Expenditure Per Student, In-state Tuition and Board costs. It is important to note that for these features their values are in the thousands hence the changes we see may seem high but when viewed as a proportion of the range they are not as drastic as we initially imagined. Based on the results of our imputation we were confident in the new replaced values and wanted to do more exploration on the various relationships that exist among the features in our data.

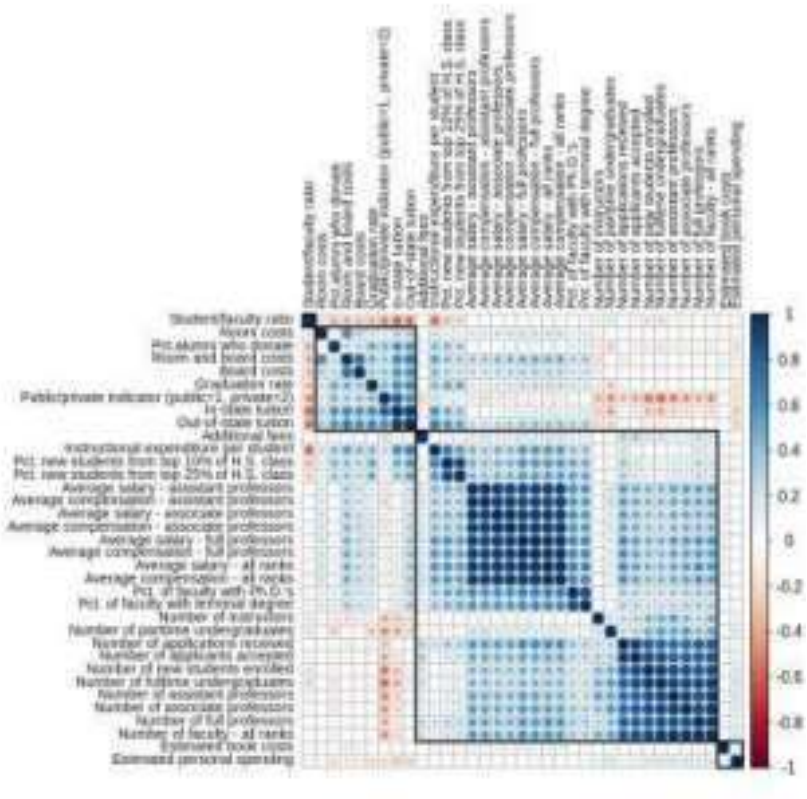
Variable Correlation

Aiming to predict the In-State Tuition makes this a Regression problem hence in our approach and model selection we will need to be conscious of using predictor variables that have a very strong correlation with our response variable as this creates the potential for Multi-collinearity.

These predictor variables being included would not add any information to the model and would potentially impact some of the coefficient values and hence the model accuracy, as such assessing feature correlation is important.

The figure below shows the correlation between all the numerical features in our dataset.

Based on the correlation plot you can tell that among some of the features there are number of variables that have very strong correlations. It is important to indicate that these are mostly strong positive correlations.



Focusing on our response variable specifically, In-state Tuition, we can see that the strongest correlations exist between itself and Out of State Tuition,

Room, Room and Board as well as

Public/ Private status of the school.

During model building we will be assessing the impact of these on the linear regression model specifically as discussed later in the report. Given the impact of the public/private it may be worthwhile to investigate I further. This led to an interesting discovery that majority of our data is related to private universities.

Approximately 700 of the universities in the dataset were private and the remaining were private.

It is commonly known that private universities tend to charge higher tuition rates than public ones and this is for a multitude of reasons some of which may be captured as features in the dataset but for the moment these can be considered assumptions.

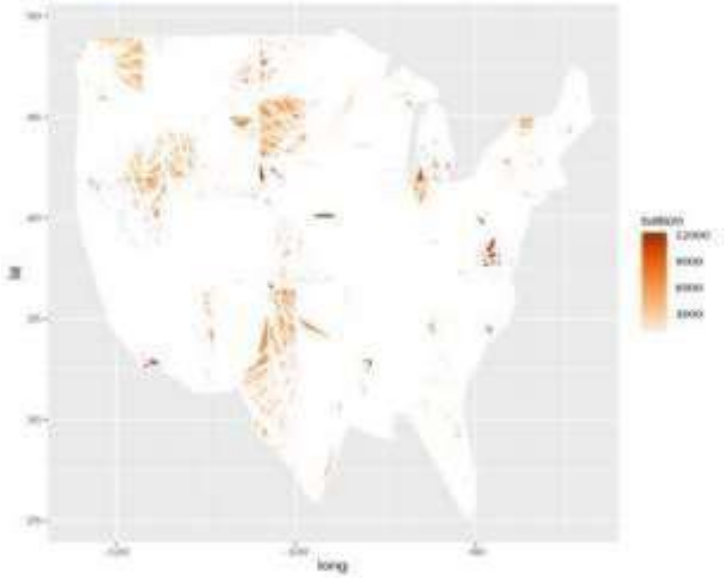
In-State Tuition By State

Part of our data includes the State location of the various universities. Each of the states in the USA operates independently of one another and pass their own policies and laws. These policies and laws are not easily quantifiable, but their impact may potentially stretch across to other areas such as the education sector which affects tuition rates and the cost of living. Cost of living impact on our dataset is covered by features in our data such as room, room and board as well as compensation and allowances paid our to faculty members.

Viewing the data as map may show be a good way to visualize the differences in tuition charged in the different states.

The figure below is a map of USA that represents the tuition charged in the states.

To determine the tuition charged by state we obtained the mean of all the universities located in that state. The tuition legend thus represents the average tuition charged by state.



One of the things we notice is the fact that the universities were clustered in the same regions of the states. Based on the university locations with high tuition rates we can tell that they are located in highly urbanized areas like large cities and are clustered in specific areas. Despite the maximum of the average tuition rate being

at $12,000, the range of the In state tuition feature is very large at $27000+ and goes down to $22,000 when we account for the outliers that may be in the In-State tuition.

For the outliers in the In-State Tuition we define them as 1.5 times added to the third quartile of the private In State Tuition. Since tuition does not go below zero there Is no need to specify a lower bound of the outliers however some of the really low charges at public institutions may be due to the discounted tuition rates from federal loans, bursaries from the government, loans from private lenders as well as balances of scholarships.

Regression Models

For our project we tried out four regression models that all made decent predictions however as we all know there is always room for improving model predictions through various steps such as cross validation, model selection, feature selection and parameter tuning that result in models that generalize the train data and create good predictions on the test data.

The models we used were the following:

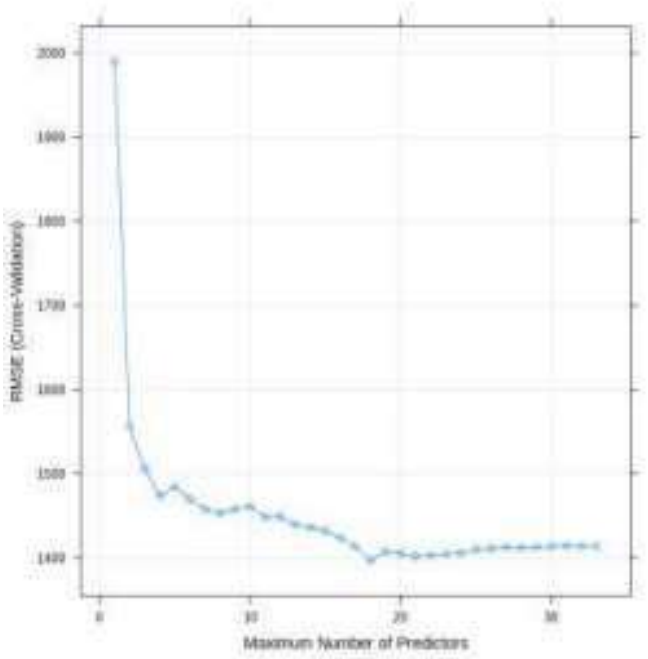
- Linear Regression Model
- LASSO Regression Model
- Random Forest Model
- Support Vector Machine Regression

In this report we will address the variables we used for these models, evaluation of the models and how the model stacks up against the others.

Linear Regression Model

Following all the data cleaning that the data underwent up until the point we began modelling the dimension of the data were as follows : 1134 rows with 35 variables.

To determine the most suitable variables for this model we used a mix of cross validation and backward stepwise selection. This method allowed us to narrow down the variables to use down to 18 variables which makes our linear model a lot less complex that it would have been had we used all the features as predictor variables.



The plot to the left represents the RMSE of the cross-validation process to determine the best number of variables to use for the model. At 18 variables we achieve the lowest RMSE of the model and thus we use these 18 variables as the predictor variables of our linear regression model.

We evaluate our models and their performance based on the RMSE, R-Squared and the Adjusted R-Squared.

For the linear mode l the above metrics

are as shown below:

Train RMSE	1340.319
Test RMSE	1500.58
R-Squared	94.04%
Adjusted R-Squared	93.91%

Given that we are trying to predict tuition where some of the values are in the tens of thousand the RMSE of the train and test data seems reasonable. The difference between the train and test

RMSE is also not very large and as such we can say the model does generalize on the train data as expected.

Looking at the Adjusted R-Squared we can agree that the model explains 93.91% of the variance that we have in the data.

Earlier on in this project we brought up the concern of the correlation between some of the feature variables. To get a better insight on if this may have impacted the model and we thus assessed the VIF of the predictor variables. We noted that six of the VIF values of the predictor values were bigger than 5 but less than 10. We will keep these variables and retain the model as is however we remain conscious of the potential impact that the multicollinearity may have on the coefficient values.

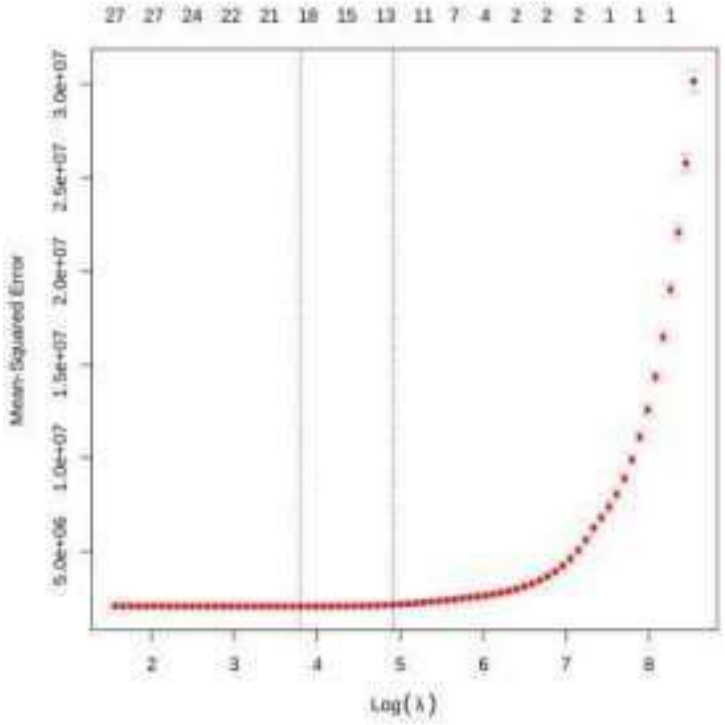
LASSO Regression Model

Our LASSO regression model was built based on the features that we selected using the backward stepwise selection.

In order to achieve an optimal model, we calculated the minimized lambda value which was returned at 18.05.

Setting our lambda to this minimized value we reduce seven out of the 18 coefficient down to zero. LASSO being a penalized form of regression makes it very good at reducing the impact of multicollinearity in the data by highly penalizing the correlated variables and in some cases reducing the coefficients down to zero.

The above statement is reinforced in this case as the features that are reduced to zero are the same predictor variables in the linear regression that are associated with high VIF and as such they do not add new information to the model.



The plot on the left shows the progression of the LASSO model on the various features.

As noted before the lowest MSE is achieved dat 13 variables and these are what we use for the LASSO model.

The MSE is scaled in the millions and as noted above the changes the between the predicted and actual values the squared values that represent the MSE are in the millions. As a result, we use the RMSE which is lower scaled thus easier to make compare against other models.

We evaluate the LASSO model on the same metrics as the linear model.

Train RMSE	1385.90
Test RMSE	1539.07
R-Squared	91.31%
Adjusted R-Squared	89.26%

The Train and Test RMSE when compared to the Linear Regression RMSE's we can tell that they are pretty much in the same ballpark. The difference is them is actually very minimal more so on the train data. The Adjusted R-Squared is slightly lower but nevertheless it explains 89.26% of the variance in the data.

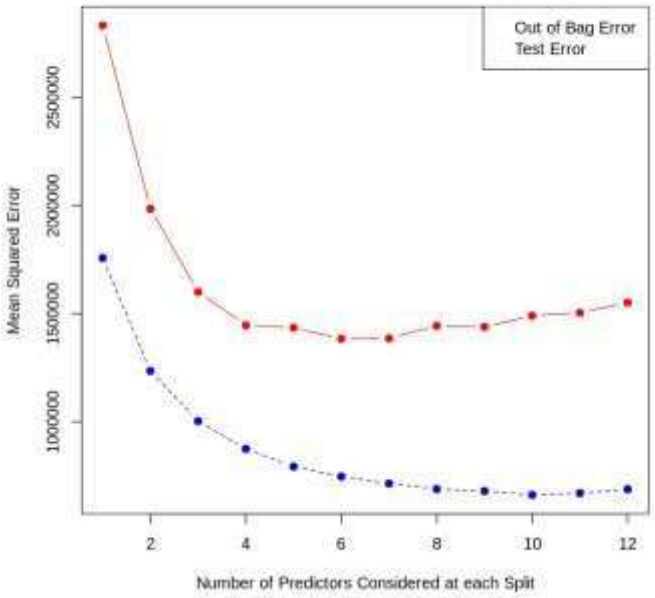
Given the minor difference between these models, I would be more in favour of the predictions made by the LASSO model due to the fact that it addresses the correlation issues that may have an effect on the linear model coefficients. Additionally, while the linear model struggles with predicting low In State Tuition we can note that the LASSO model seems just as accurate on lower value in state tuition as it is with high value predictions of In state Tuition.

Random Forest Model

Random Forest is one of the most versatile models and is very good for model selection as well as use in regression modelling.

For our random forest, the lowest number of predictors that we considered at each split was 8.

Below is the Out of Bag Error and the Test Error that was obtained when we tested the random forest model.



We notice that the Out-of-bag error is very high. The Out -of- bag error is a lot higher than the test error. As we stated earlier at 8 features at each split, by increasing the number of features at each split we may be able to address the large out-of-bag error.

Model Evaluation:

Train RMSE	453.766

Test RMSE	843.175
R-Squared	77.22%
Adjusted R-Squared	97.14%

Despite the RMSE values that we have being very low we can see that there is a very large difference between the train and test RMSE is large. Additionally, the model has a very low Adjusted R-squared compared to other models and hence performs poorly on the test set.

Support Vector Machine Regression Model

Traditionally Support Vector Machin models have been heavily used in solving classification problems however their use also stretches into the realm of regression and as such it can be applied to our in-state tuition prediction problem.

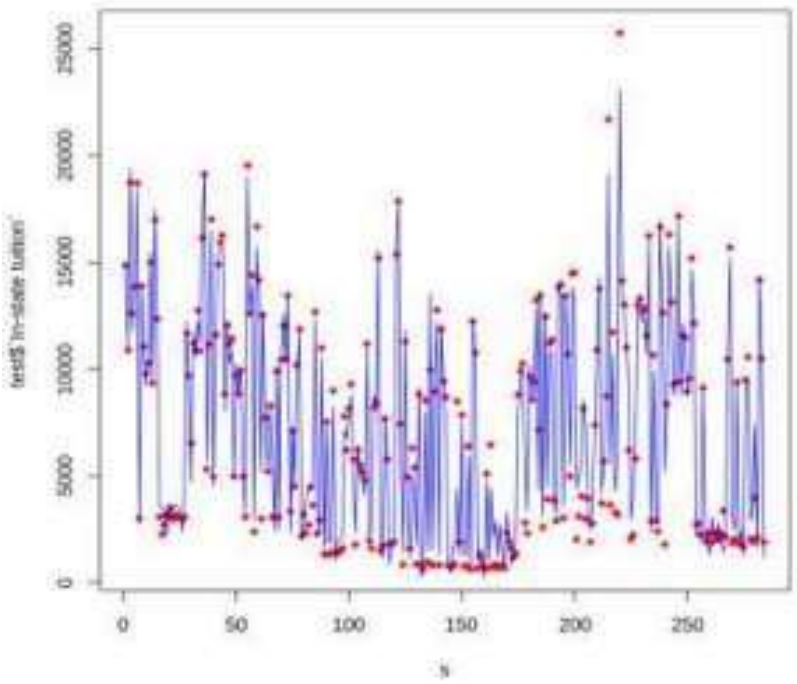
In our case the tuned hyper parameters were:

- Kernel set to radial.
- Hyper plane set to feature space.

Train RMSE	809.161
Test RMSE	964.561
R-Squared	96.58%
Adjusted R-Squared	94.33%

All factors considered despite having a higher train and test RMSE than the Random Forest, this model is better as it addresses the pit falls that random forest made. The difference between the

train and test RMSE in SVM is a lot lower and the Adjusted R-Squared is the highest we have observed.



On the left we have a visual representation of the Support Vector Machine model that shows the actual values inn red and the predicted values in blue. For the points that do not have clusters around them we can see that the predictions do really approximate the actual values.

FINAL REMARKS:

Based on the RMSE's we can say the generalization of this SVM model for the test set is a lot better and there is potential for the model to be more efficient by tuning the hyper parameters to fit the data better. We could standardize or normalize all the data across the entire dataset which may result in improved model accuracy.