Comprehensive Analysis on Rice Harvest Price in Andhra Pradesh, a state in India

Santosh Reddy Edulapalle, A20501739, *sedulapalle@hawk.iit.edu*
Venkata Siva Rupesh Akurati, A20501754, *vakurati@hawk.iit.edu*
Jobin Joyson, A20419597, *jjoyson1@hawk.iit.edu*

Abstract

The overall goal of this research surrounded the concept of understanding the popularity of rice production in India and how the state of Andhra Pradesh can provide affordable prices for rice. The datasets used for this research are from the International Crops Research Institute for the Semi-Arid Tropics(ICRISAT). After analyzing the data and training an appropriate model, we found that the major influences on the price of rice was from the years, average male wages, and the precipitation during the Kharif season

Overview

Crop production has always been an essential part of the world. Maintaining these production systems can serve as a challenge because despite the continuous efforts from many regions, the world still has millions suffering from starvation. In order for regions to provide for the hungry, the goal of crop producers should be to make products affordable to the general population. Analyzing the effects on crop prices will allow us to see what must be done to maintain an ideal price.

For our research, our target crop was rice and we focused on India as it is the second largest crop producing country in the world. We used the datasets provided from International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). The original amount of datasets provided was massive, so to be able to do an accurate analysis with our given timeline we decided to focus on the state of Andhra Pradesh which is ranked among the top five states in India for rice production.

As for our target variable, we chose to look at how the price of rice is affected through the factors involved in rice production. Specifically we looked at farm harvest price, which is the price of the crop at harvest time. Wholesale prices are subject to a 15-20% increase from these prices. This will give us an under-

standing of what systems in rice production alone influence the price of rice. By identifying these variables, we can make an analysis of how rice producers in Andhra Pradesh can maintain a reasonable price.

With our target variable being qualitative we approached a linear regression model on our training dataset. From inspection on the regression metrics, we used methods from the olsrr package in R to perform an AIC forward model selection process. This allowed us to compare many linear regression models and find the model that contained predictors with the least AIC value and overall shrink the model from 77 predictors to 34 meaningful predictors. Further analysis was performed to the original linear regression model to see if we can get a more accurate model version of the model through variable selection. We used LASSO methodology to shrink the coefficients and bring the coefficients of the non-significant predictors to zero. Both models performed well on the testing dataset and clear conclusions were drawn from the results.

Data Processing

Our first challenge was to manage the vast amount of data spread across 11 datasets, despite focusing on rice in Andhra Pradesh. In terms of year, we removed observations before 1990 because only the environment dataset contained information before 1990. As for the end range, we limited observations until 2015 due to the majority of datasets ending observations by this year and only a couple having data until 2017.

Deleting the data with years ≤ 1989 & ≥ 2016

Once we got a date range from 1990 to 2015, we noticed that there were still more missing values in some datasets within the range. Since Andhra Pradesh has 13 districts, we took the averages in each of these districts to fill up the missing values. These included: Irrigation Area, Rice Farm Harvest Price, Input Wages, and Season Fertilizer Consumption. In order to reduce variance that may occur as time passes, such as environmental changes, the averages were calculated within 3 intervals: 1990-2000, 2000-2010, 2010-2015.

Handling missing values

- We first glanced over all datasets and removed any records with a value of -1. We were able to see a number of NAs later after integrating all the datasets into a single df.all dataframe. Following a comprehensive analysis, we discovered that since a few observations in a few datasets had been destroyed, those years' observations were causing NA when merging with other datasets. Therefore, we handled the largest amount of missing data, df.input.wages, by treating the appropriate years. Additionally, for a few tables, we lacked district codes 503 and 504, therefore we acquired fresh and updated information.

- Then, we handle those specific columns by executing the mean of each district in that particular year gap (10 year average for missing districts using filter()) as we take summaries of all tables and verify if any table has minimum value -1.
- After performing steps 1 and 2 and joining the tables, we still received a small number of NAs. However, after reviewing the summary of each join, we discovered that the 1995 district 503,504 has two missing values in the df.rice.farm.price table. and this wasn't given a -1 code. Instead, it is entirely absent. identical situation for all districts' df.input.wages for the years 2010, 2011.

  The NA values were handled for the below dataframes:

- df.irrigation.sourcewise.irrigated.area
- df.rice.farm.harvest.price
- df.input.wages
- df.inputs.season.fertilizer.consumption

Merging Redundant columns

Most of the datasets had collected information in terms of months, which resulted in quite a few redundant columns. For example, rainfall was calculated monthly so there were 12 columns for rainfall alone. We decided to split the months by agricultural season in India: Kharif, Rabi, and Zaid. With these three new columns we could take an average of the months that each season covers. Kharif would take the average from July to October, Rabi would take average from November to February and Zaid would take average from March to June. We also got rid of the state name column, since we are focusing on only Andhra Pradesh, as well as the district name column since we have district ID as a reference.

Merging all dataframes

Now that we have completed our datasets, we will attempt to integrate them into a single data frame. Here, we'll use the joins function from the dplyr package. Because we don't want to lose any data, we will utilize a full join. The missing values will be dealt with later, as planned. Due to the cartesian product, we chose to go on and join on Year while focusing on the fresh data first because ee are getting a lot of rows.

Normalization: Standard Scaling method

Different datasets had various numerical ranges such as 0-100 or 10,000-100,000. We scaled these values to lower the variance in our analysis. With these modifications we were finally able to join all the datasets to one dataframe to prepare for our analysis.

$z = (x - \bar{x})/s$ for scaling
and, $z = (x * s) + \bar{x}$ for re-scaling back.
where, $\bar{x}$ is the mean of x and s is the standard deviation of x

Factorizing the data

To have better interpretation of the results, we have factored the data on 'district code' variable.

Data Analysis

To get an idea of the data we have before any feature engineering is done, we looked at changes our target variable, rice harvest price, had over the years.

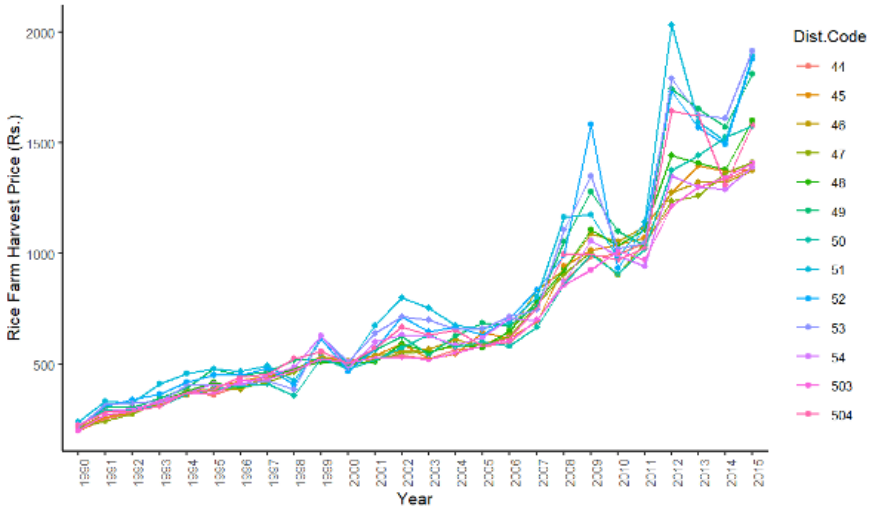

Figure 1: Rice FHP in Andhra Pradesh by Year and Dictrict

We can see that the harvest price of rice has steadily increased across the years in all districts. In the early 2010s there were more erratic changes in the price as it continued to increase.

Before we go into modeling to get the details of what influenced this change, we can make some assumptions on the data. Since most of our data is linked to factors that affect the production of rice, we assumed that this increase in rice price may be due to a shortage in production of rice. To investigate this, we visualized the changes in rice production and rice yield across the years in all the districts of Andhra Pradesh:

Initially it does seem that the production has changed drastically over the years, especially in the early 2000s. However, there are equivalent decreases in produc-

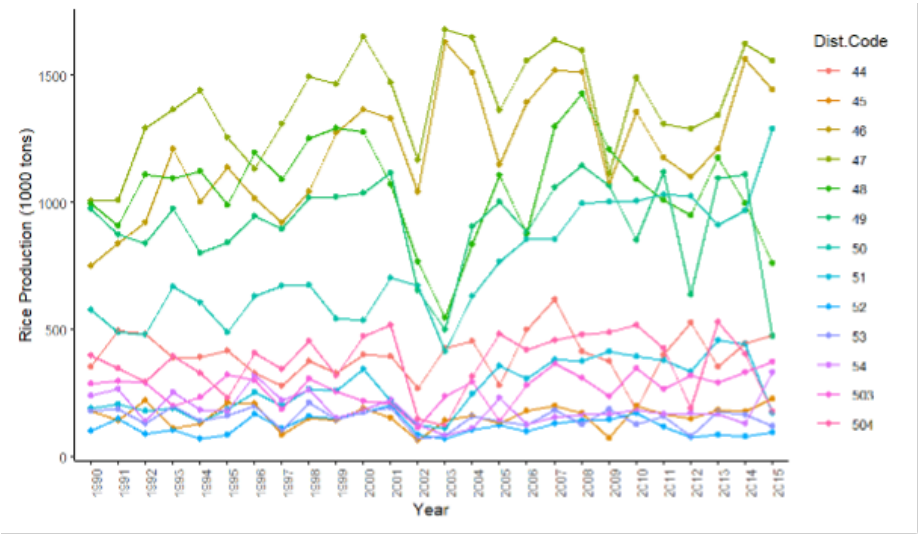

Figure 2: Rice production in Andhra Pradesh by Year and District

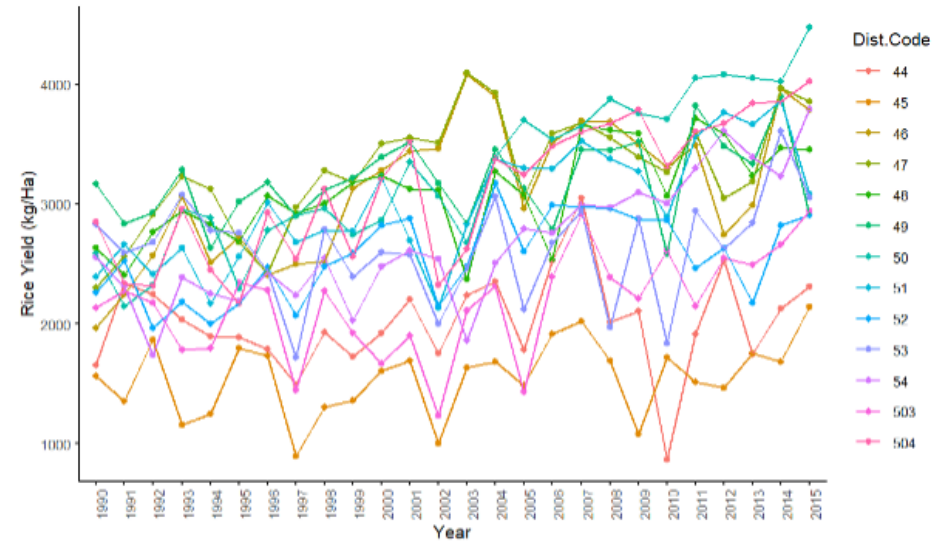

Figure 3: Rice yield in Andhra Pradesh by Year and District

tion following the initial increases and vice versa across the years. For example, looking at district 48, there was a major decrease in production from 2000-2003. But production picks up again, back to almost the same amount a few years later from 2003-2007. So overall there hasn't been definitive change in the production.

We can also see that it is similar in the case of rice yield throughout the years. Despite the many ups and downs within every 2 years, the rice yield in each district stayed relatively the same from 1990 to 2015.

This could mean two things: the increase in price over the years could be due to inflation as we had originally assumed, or there were other factors that played a more significant role in increasing the prices. Judging by the data we have at hand, most variables play a role in the production of rice (Rice Area, Canal/Wells/Tube Area, Consumption, Rainfall, etc.). The possible predictors that can influence price greatly outside of rice production would be the wages for each worker. We can visualize the wages for the workers throughout the years and see what we can interpret from that:

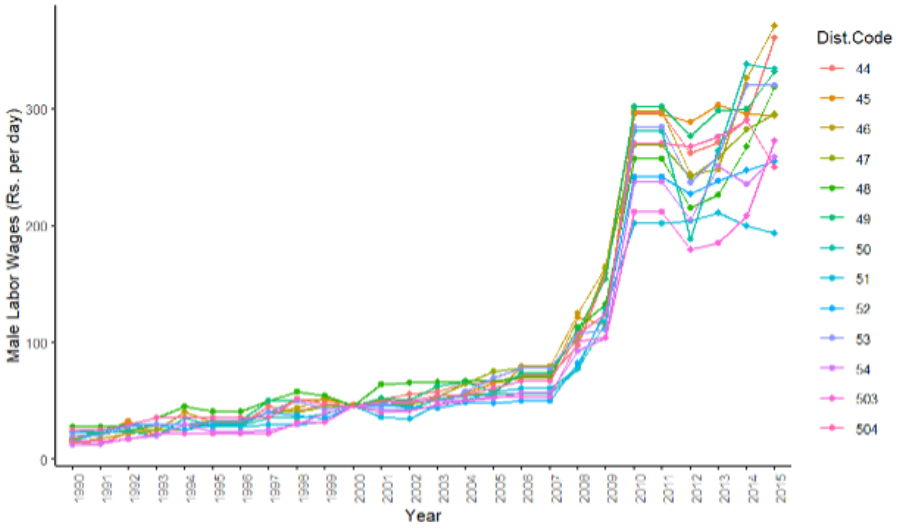

Figure 4: Male Labor Wages by Year and District

As we can see from the visuals, the wages for the workers have increased steadily until 2007. From 2008-2015 the wages have increased rapidly. This reflects what we saw with the price of harvest as well. This makes intuitive sense because as the cost of workers goes up, farmers' charge on rice harvest price will reflect that. This further supports our initial assumption of inflation playing a key factor in the increase of rice harvest prices. Now that we have a general idea of the data, we can go into the statistics of the data and see what features play a role in price that we cannot easily see.

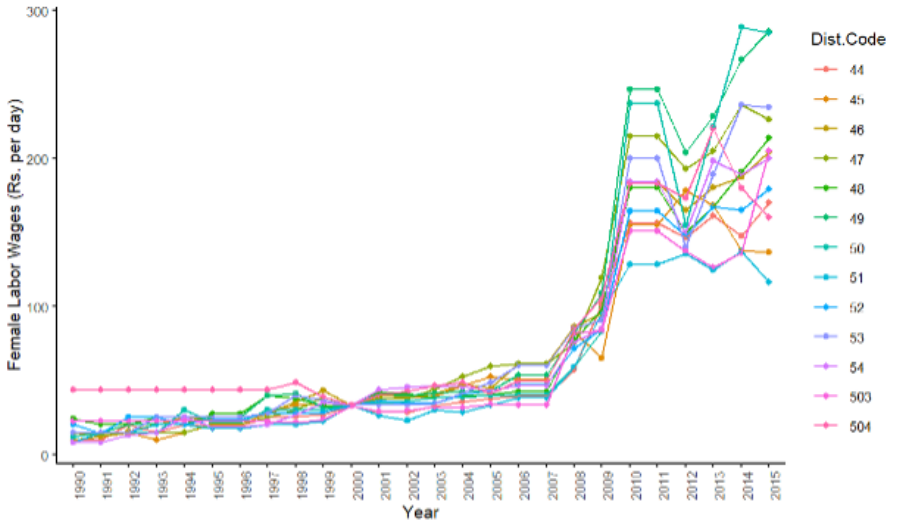

Figure 5: Female Labor Wages by Year and District

Modeling

Linear Regression

Unlike the conventional method of splitting the data in 80-20 fashion, we followed another approach. Since the significance of this project stresses on the historical data and how well it can be modeled to predict the future statistics, we have considered the data from 1990 to 2010 for the training and the data from 2011 to 2015 has been considered for testing.

The initial implementation of Linear Regression by considering all of the available predictor variables was performed and results of the summary of the model are shown below.

```
##
## Residual standard error: 0.4909 on 193 degrees of freedom
## Multiple R-squared:  0.9498, Adjusted R-squared:  0.9293
## F-statistic: 46.23 on 79 and 193 DF,  p-value: < 2.2e-16
```

Figure 6: The results of the model with all predictor variables

The adjusted R-squared obtained by considering all the predictor variables is 0.9293.

AIC Model

When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; so some information will be lost by using the model to represent the process. Since AIC (Akaike information criterion) estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model, so we implemented AIC with forward prediction as this will help us to prevent overfitting. Thus, AIC provides a means for model selection.

In forward selection, the first variable selected for an entry into the constructed model is the one with the largest correlation with the dependent variable. Once the variable has been selected, it is evaluated on the basis of certain criteria. The overall predictors chosen for further consideration in the project has been displayed below by the model as below:

```
##
##                                    Selection Summary
## ------------------------------------------------------------------------------------------------
## Variable                                    AIC      Sum Sq       RSS      R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------------------------------
## Year.x                                   640.037    763.425   163.053   0.82401       0.82336
## STATE.MALE.AVERAGE.FIELD.LABOUR..Rs.per.Day.  592.788    790.340   136.139   0.85306       0.85197
## Kharif.percipitation.mm                  548.044    811.763   114.715   0.87618       0.87480
## PERMANENT.PASTURES.AREA..1000.ha.        534.235    818.218   108.261   0.88315       0.88140
## TOTAL.KHARIF.CONSUMPTION..tons.          525.246    822.489   103.990   0.88776       0.88566
## CURRENT.FALLOW.AREA..1000.ha.            515.619    826.825    99.654   0.89244       0.89001
## NET.CROPPED.AREA..1000.ha.               506.745    830.716    95.762   0.89664       0.89391
## Kharif.evapotranspiration.actual.mm      501.405    833.257    93.222   0.89938       0.89633
## Rabi.rainfall.mm                         499.017    834.743    91.735   0.90098       0.89760
## OTHER.WELLS.AREA..1000.ha.               495.324    836.636    89.842   0.90303       0.89933
## LAND.PUT.TO.NONAGRICULTURAL.USE.AREA..1000.ha.  493.804    837.787    88.691   0.90427       0.90024
## Rabi.evapotranspiration.potential.mm     481.879    842.198    84.281   0.90903       0.90483
## Dist.Code                                450.985    857.549    68.929   0.92560       0.91840
## NITROGEN.KHARIF.CONSUMPTION..tons.       444.920    859.556    66.923   0.92777       0.92046
## Zaid.rainfall.mm                         441.380    860.900    65.578   0.92922       0.92174
## CULTIVABLE.WASTE.LAND.AREA..1000.ha.     439.266    861.881    64.597   0.93028       0.92259
## Zaid.temp.max.c                          436.999    862.883    63.596   0.93136       0.92348
## TANKS.AREA..1000.ha.                     436.029    863.571    62.907   0.93210       0.92400
## Rabi.percipitation.mm                    435.113    864.239    62.239   0.93282       0.92449
## Rabi.evapotranspiration.actual.mm        431.318    865.547    60.932   0.93423       0.92577
## Rabi.temp.max.c                          430.942    866.015    60.464   0.93485       0.92611
## Zaid.evapotranspiration.potential.mm     430.759    866.556    59.923   0.93532       0.92639
## Zaid.percipitation.mm                    430.267    867.100    59.378   0.93591       0.92675
## ------------------------------------------------------------------------------------------------
```

Figure 7: Final predictor variables chosen by the model

```
##
## Residual standard error: 0.4995 on 238 degrees of freedom
## Multiple R-squared:  0.9359, Adjusted R-squared:  0.9268
## F-statistic: 102.2 on 34 and 238 DF,  p-value: < 2.2e-16
```

Figure 8: Results of the AIC model

The AIC model (refer figure 8) produced an adjusted R-squared value almost similar to that of linear regression model but with only 34 predictors which is considered to be a significant improvement in the modification of the model.

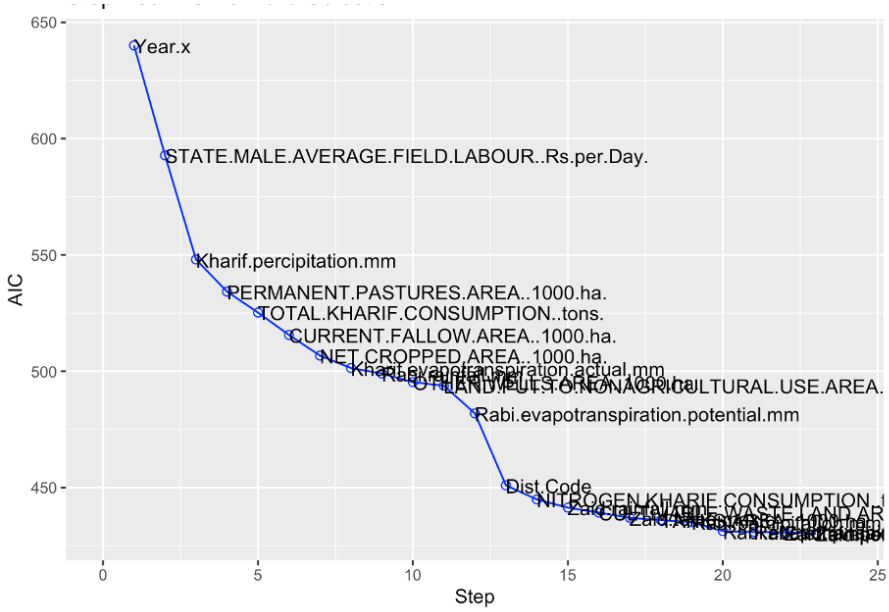

Figure 9: Graph of AIC forward selection model showing the predictors

From figure 9, the AIC forward selection model shows the predictor 'year' has been the most influencing variable followed by 'male wages', 'kharif precipitation' etc.

The model characteristics shown in figure 10 depicts two pictures in the top row which represent the residuals being not too deviated from the zero values showing a decent fit. While the top right picture shows the residuals being too close to the linear dotted line which also represent the good fit of the model.

Lasso Regression

To enhance the prediction accuracy and interpretability of the resulting model, we also performed LASSO regression by performing both variable selection and regularization.

we perform cross validation to find the best lambda that gives the minimum test MSE and use this `best_lamda` to train the Lasso model.

The coefficients have been reduced by our Lasso Regularization in comparison to the linear fit, and the majority of them have been reduced to virtually zero, while for many of the predictors, no coefficients were produced by the lasso.

In the figure 12, we can see the dashed lines are the $log(\lambda)$ values corresponding to the λ_{min}(left dashed line) and λ_{1se}(right dashed line). λ_{min} is the value for which the model has the lowest cross-validation mean squared error. This error

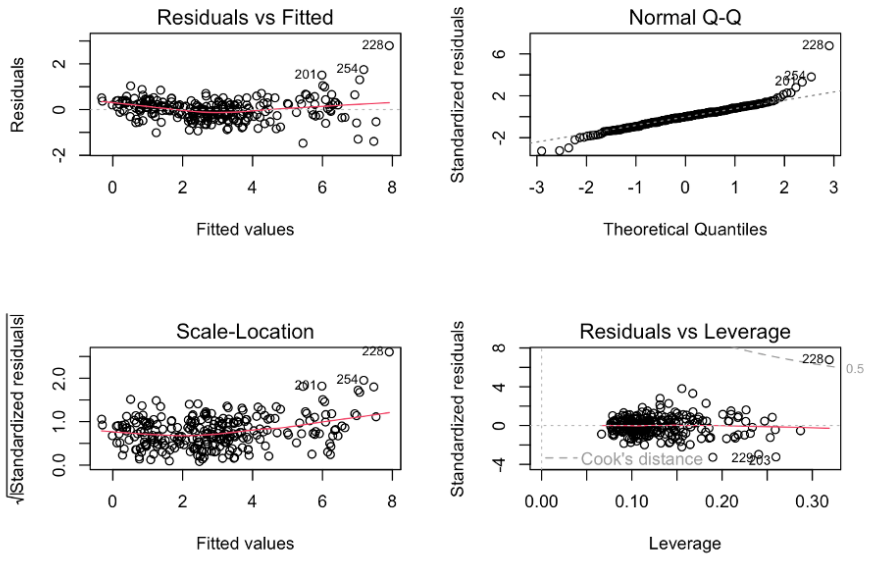

Figure 10: Characteristics of AIC model

has a certain variance/standard error (depicted by the gray whiskers to every red point which you cannot see properly in the left part of the plot, but clearly in the right part). λ_{1se} is a value that is a bit less prone to overfitting, we get it if we extend the height of the upper whisker of the min dot to the right until you reach the last point which is still below that imaginary line (thus, within one standard error of the minimum value).

The numbers shown on top in figure 12 are the number of non-zero regression coefficients in the model (thus, the number of included features). From left to right along the x-axis, with increasing λ we have fewer variables in the model, since the penalty for inclusion of features is weighted more heavily.

Results

To interpret the results, we de-standardized the data using the rescaling function and we now split data again for rescaled data and will test our model on rescaled observations because here we are only splitting the data based on year and it is not a random shuffle. We then tested our model of 'All predictor' and we can see this model was unable to accurately forecast the outcomes as RMSE is 3076. The plots of Linear Regression model along with other tests are also produced below.

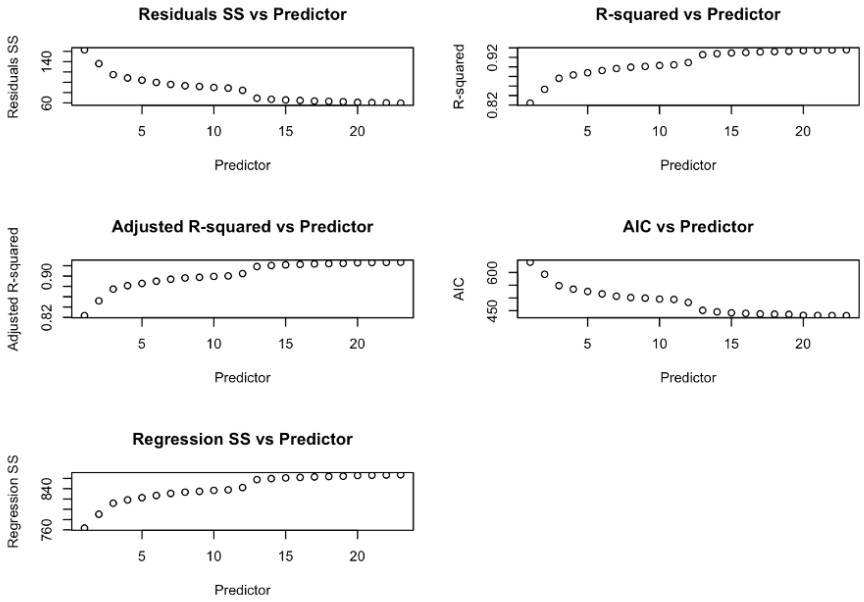

Figure 11: Plots of various residuals of the AIC model

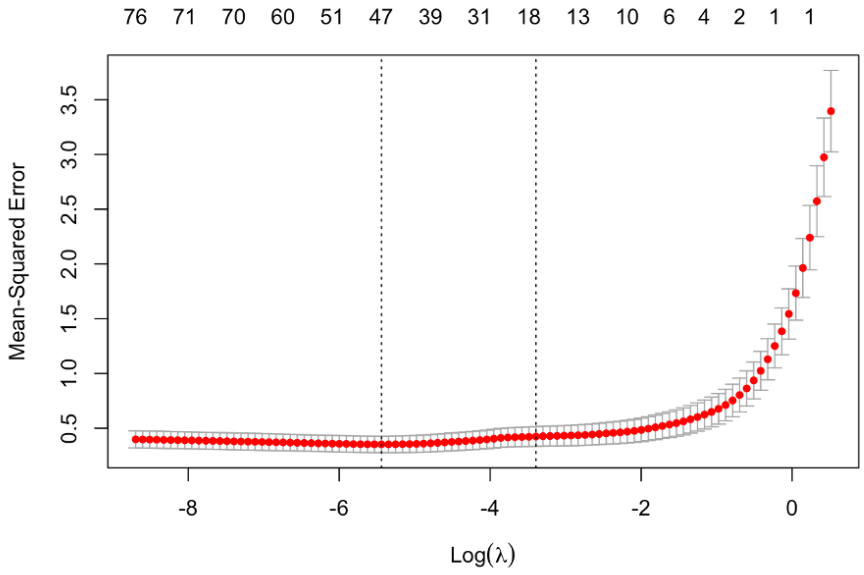

Figure 12: Plot of Test MSE vs λ

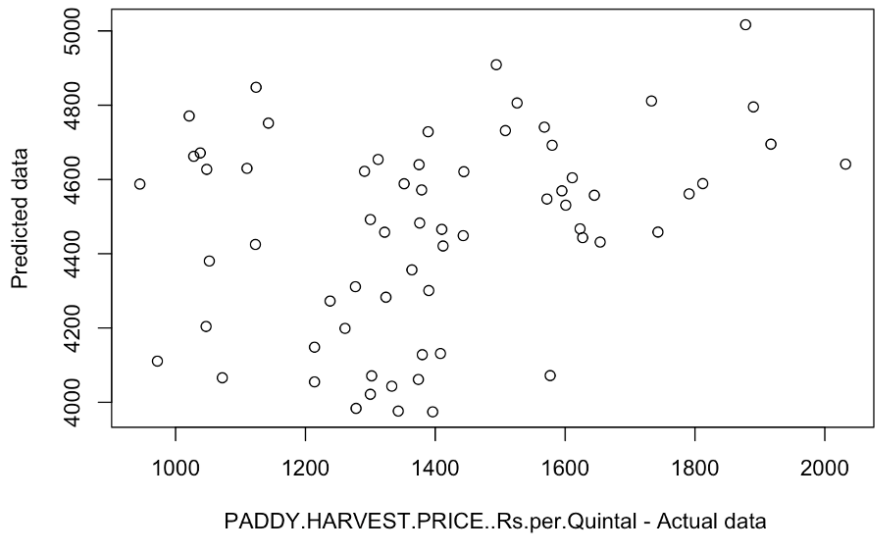

Figure 13: Linear Regression Model with all predictors: Actual vs Predicted values of Test data

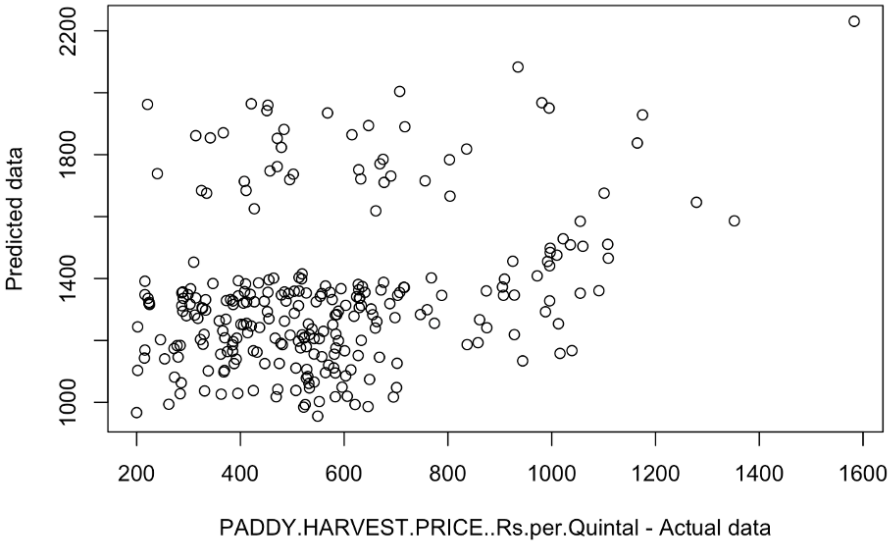

Figure 14: AIC model: Actual vs predicted values of Train data

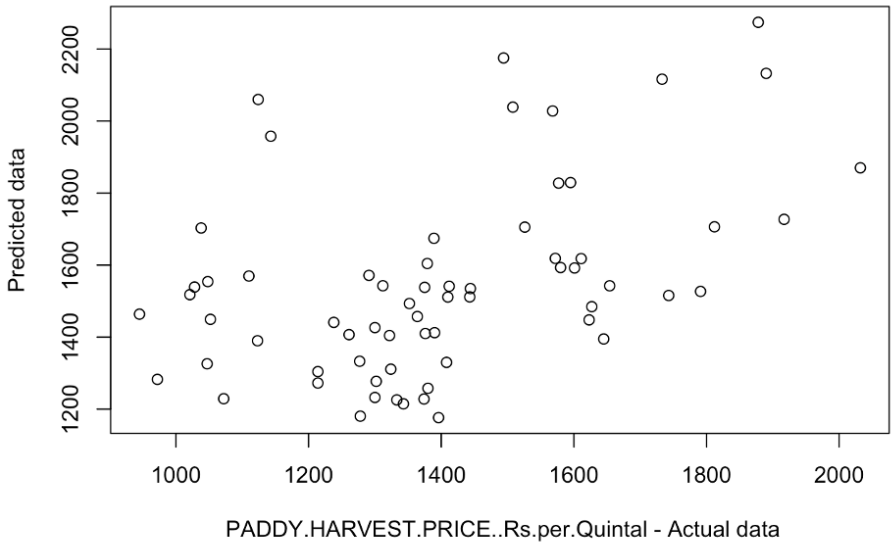

Figure 15: AIC model: Actual vs predicted values of Test data

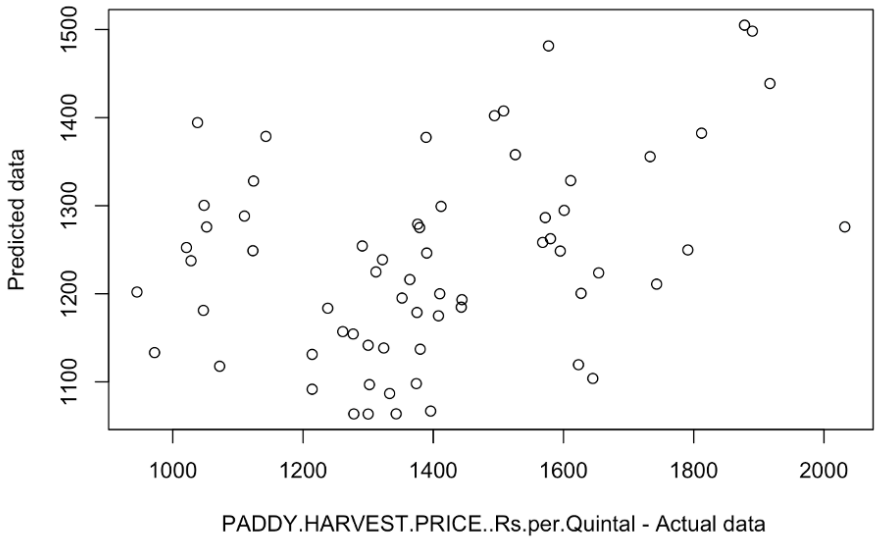

Figure 16: Lasso model: Actual vs predicted values of Test data

Conclusion

We used our training dataset to approach a linear regression model with our objective variable being qualitative. We performed an AIC forward model selection procedure using techniques from the olsrr package in R based on examination of the regression metrics. With the models implemented, we were able to minimize the number of parameters from 137 to 77 after thoroughly cleaning, analyzing, and transforming our data. We also used feature engineering to reduce the number of reductant rows that were generated when the dataframes were joined. For observations of size 338, even 77 characteristics is a fairly significant number. By building a model with all 77 parameters, we verified this, and the test RMSE was 3076. We chose AIC and Lasso to compare the outcomes after deciding to further decrease the features using either of the various feature selection algorithms. Our test RMSEs for the AIC and Lasso models are 304 and 281, respectively. Compared to the model with all predictors, we can observe that both the models accurately anticipated the results.

Data Sources

Data files: ICRISAT-District Level Data

Terminology and their units of measurements: Terminology and UOM

Source code

Please visit our GitHub page here to access all relevant content of this project.

References

Chakraborty, Ananya Murray, Emmanuel. (2011). Rice Production Productivity in Andhra Pradesh. 10.13140/RG.2.1.2919.1203.