Type of forest cover prediction

Prabhu Chaitanya cheedalla

Pranav Samudrala

ADTA -5130

Introduction

Forest cover is an area of land that is covered by forests. Forest covers are differentiated based on the conditions that the land is suitable for growing a particular type of vegetation and the weather conditions. There are a lot many commercially viable trees that grow naturally in forests like pinewood(used for furniture), aspen(medicine), fir, willow, etc..,. The growth of these trees is affected by factors like the elevation of the tree from land(like a mountain), the slope of the land, soil type, distance from the water source, sunlight intensity, etc..,

We know the implications of the decreasing forest covers, analysis of factors that affect forest covers helps to maintain the conditions that forest required the most and helps in maintaining the forest cover.

There are a lot many timber industries trying to figure out the best way to grow these trees rather than cutting the forests. Though the timber industries have their plantations, the output is not so quick and good, so they ultimately rely on forests for more wood which grows naturally without human intervention. The problem with their plantations is the land they choose, predicting the best land suitable to grow the specific tree solves the problem by giving more yield with less human intervention.

We can achieve this by the data of the forest types according to the conditions in the forest.

Hypothesis

1. Forest covers as we know are grown mainly in wilderness areas that are far from roadways, have adequate water, light, and rainfall as a result of their elevation from the land. There were some proposed ideas for man-made forests, which can be grown in human habitats by maintaining sufficient conditions as in the forests, there continues to be debated that this cannot be done. It is hypothesized that type of forest cover depends on certain conditions that man cannot create like the land elevation, its distance to water source, required intensity of light, and If there is a requirement of growing specific trees which resemble a type of forest cover, these factors should be taken in to account.

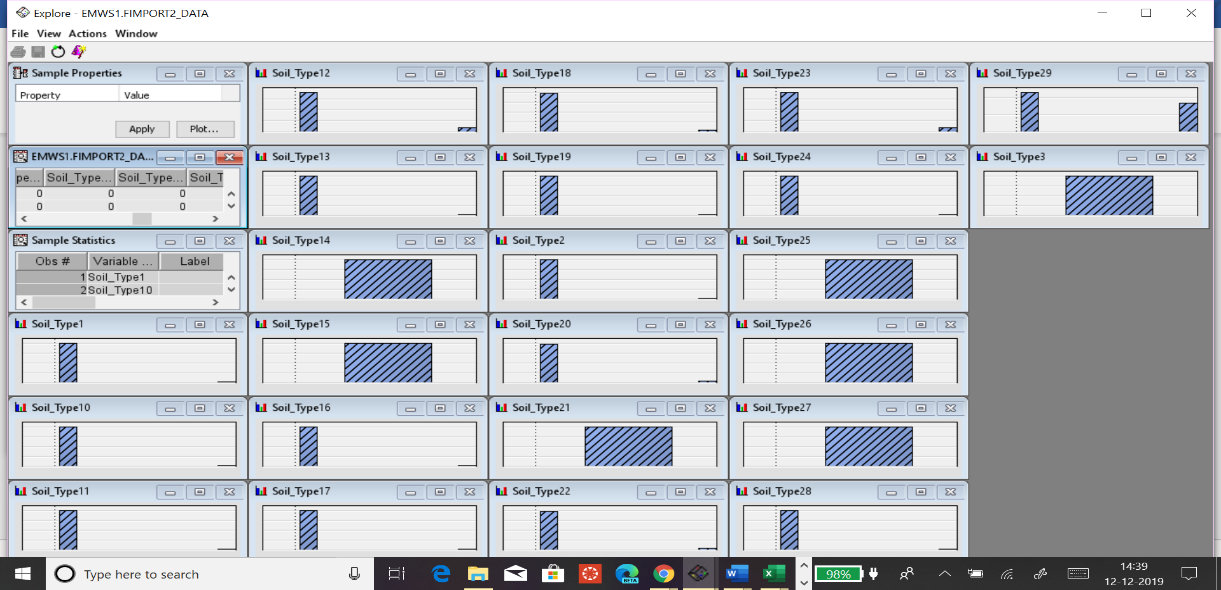
2. It is hypothesized that Spruce/Fir forests are grown in elevated regions above 1700 meters and the chance of growing these trees below 1700 meters is low. This is to analysed and proved.

Data pre-processing

The dataset has cartographic data collected in 1998 from four wilderness areas located in the Roosevelt National Forest of northern Colorado. The dataset consists of specifications of the following seven forest covers: Spruce/Fir, Lodgepole, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, Krummholz.

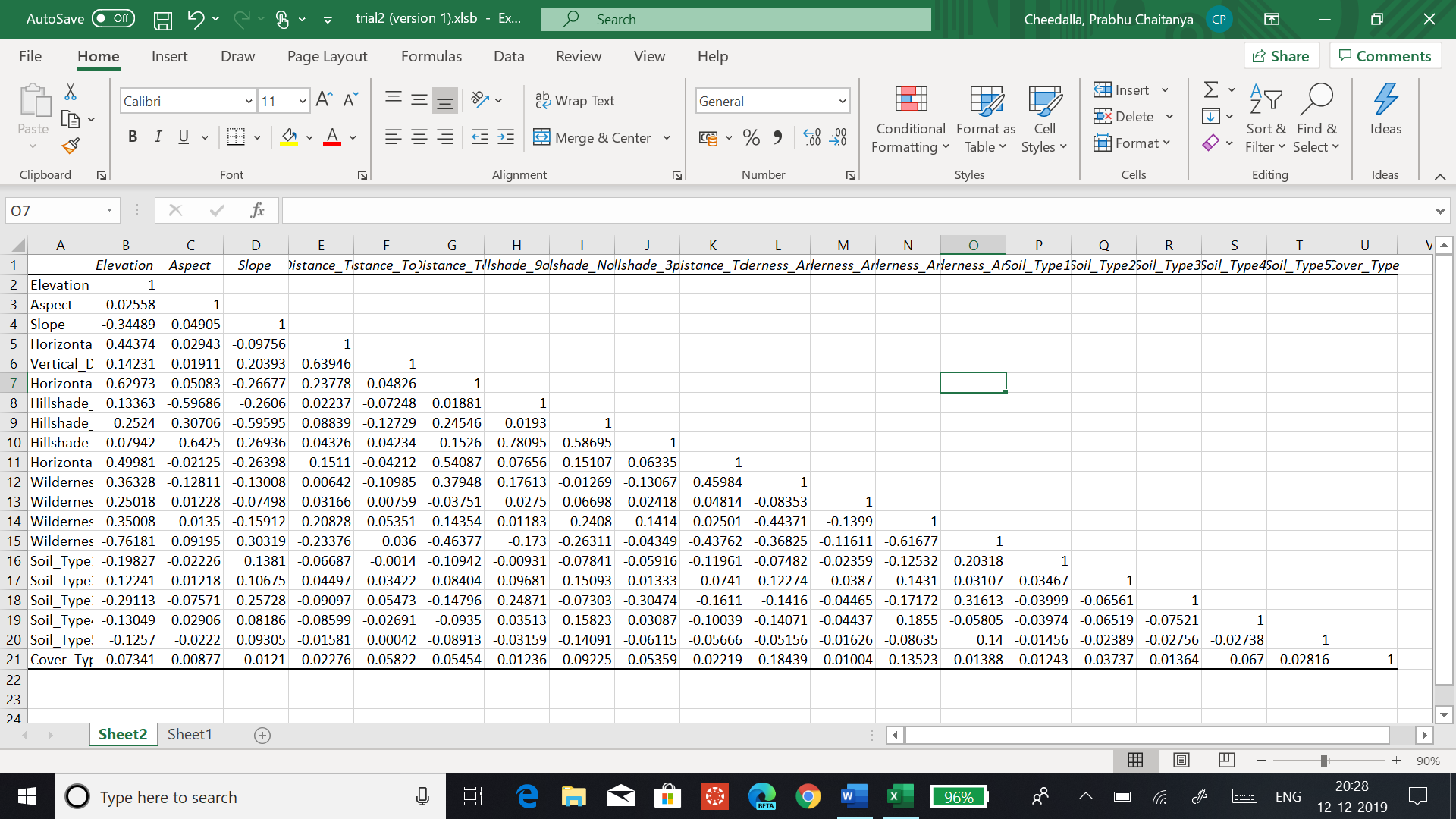
We start the analysis by pre-processing:

1. Dataset has the total instance of 581012 with the distribution of forest cover levels in percentages of 36.4,48,6,0.4,1.6,2.9,3.5 as the data is unevenly distributed, we perform stratified random sampling for the data by taking an equal number of random values from different stratum( here stratum being each forest cover type). The result obtained is 1208 observations with 15% of each forest cover type.
2. We remove the soil\_types 6 to 40 as they are not as important, also have no observations in the data as in the below diagram(mostly one-sided), and also to decrease the complexity of execution.



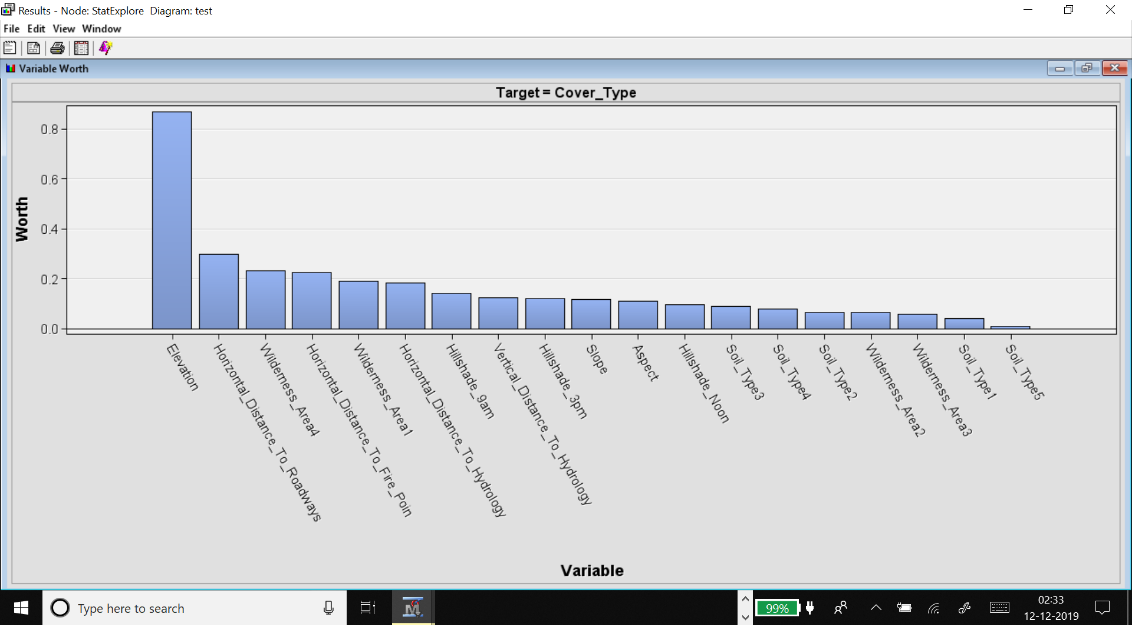
Soil type graphs

To this data, we constructed a correlation matrix, which shows that the wilderness Area1 and forest cover type are strongly correlated with magnitude ( 0.18) followed by elevation with (0.09) and weakly correlated with Hillshade\_9am(0.018).



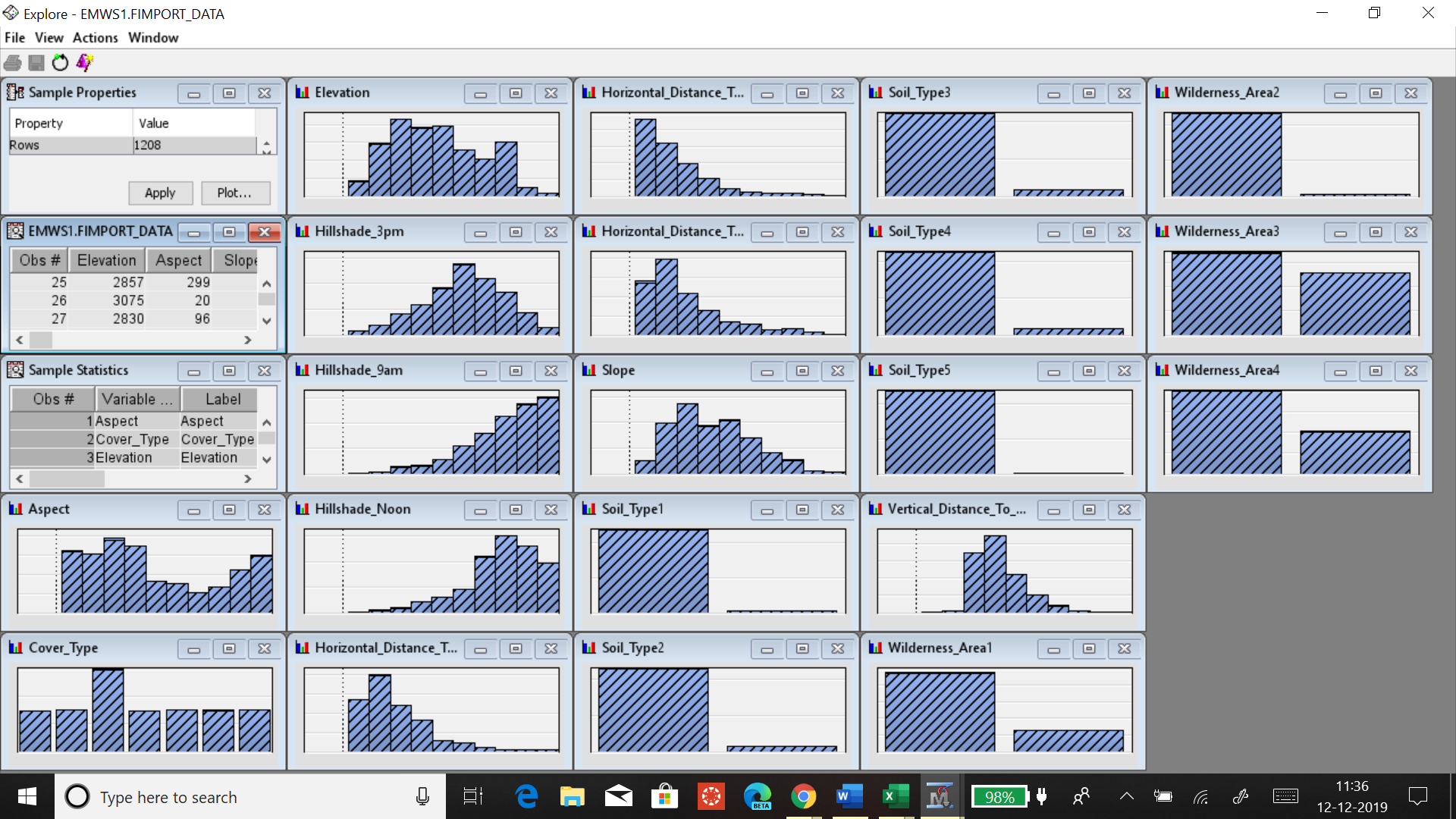
Analysis

Bar graph showing the variable worth with respect to the target



The above graph shows the variables arranged according to their importance to estimate the target variable(forest cover). This is performed in Sas enterprise miner based on the entropy value.

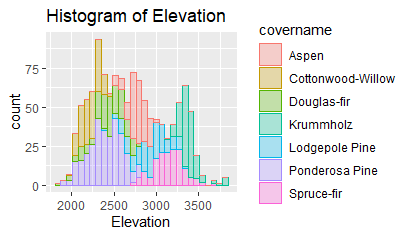
Entropy : -∑ pi log pi , pi is the probability of expectation of respective forest type.



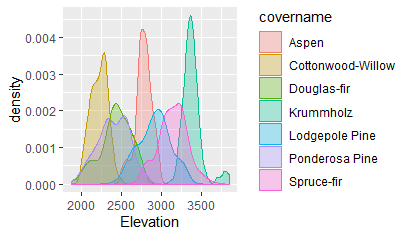
The above chart shows, the histogram distributions for all the variables, it can be observed that soil type has many one-sided observations, Horizontal distance to the roadway is positively skewed Hillshade\_Noon is negatively skewed, while Hillshade\_3pm has a normal distribution.

We perform operations to the target variable and variables that are more correlated with the target variable and more worth factor with less skewness.

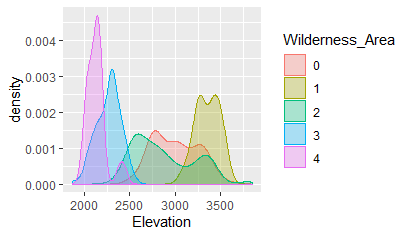
Histogram of elevation according to the forest cover type:



The above histogram performed in R by using ggplot, shows the pattern of the forest cover type with the histogram, there is a class separation observed, hence it is best used for regression.

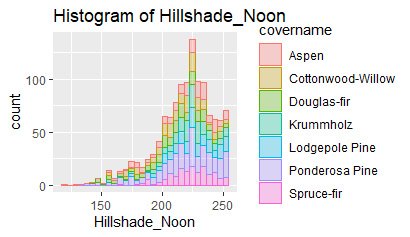


This graph shows a density plot of elevation vs forest type.



This graph shows the density of the wilderness area with elevation.

From the above two plots, it is verified that the correlation between the wilderness area, elevation with the forest type is strong, as there is a class separation clearly visible.



The above graph shows that HillShade\_noon is negatively skewed even though it shows the class separation of forest type.

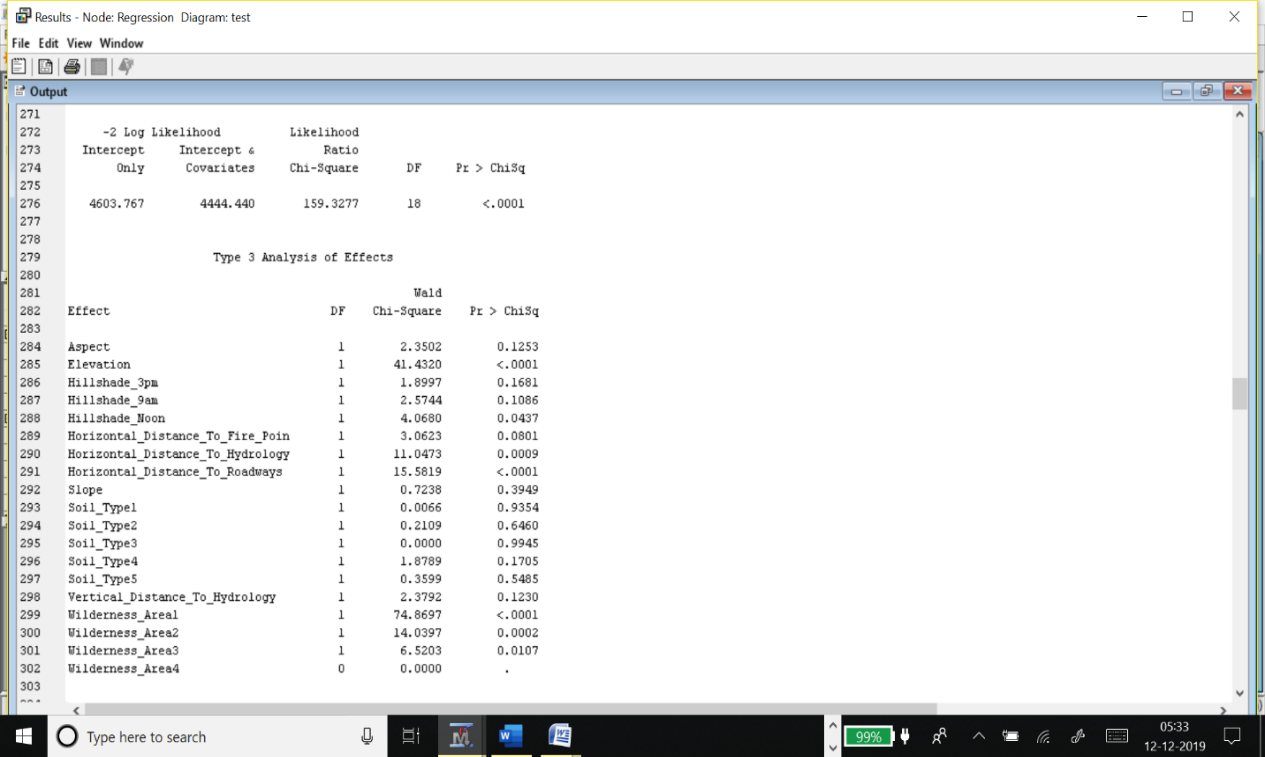
Probability of forest type prediction

The target variable for the dataset is ordinal with seven levels(seven forest types), linear regression cannot be used here, because linear regression cannot fit the binary or ordinal data, and it cannot give the output in probabilities i.e.., between 0 or 1.

Logistic regression

We perform logistic regression for forest\_cover\_type as target variable.

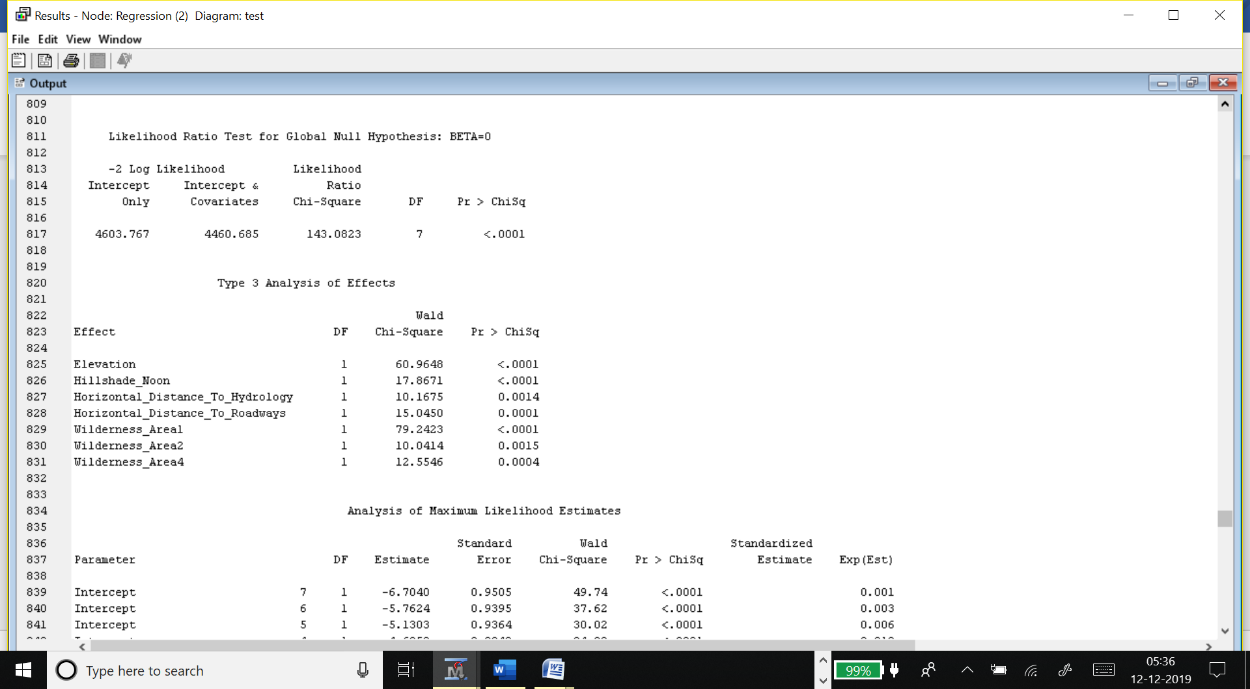
Logistic regression finds the logit of odds, from which we can derive the probability.



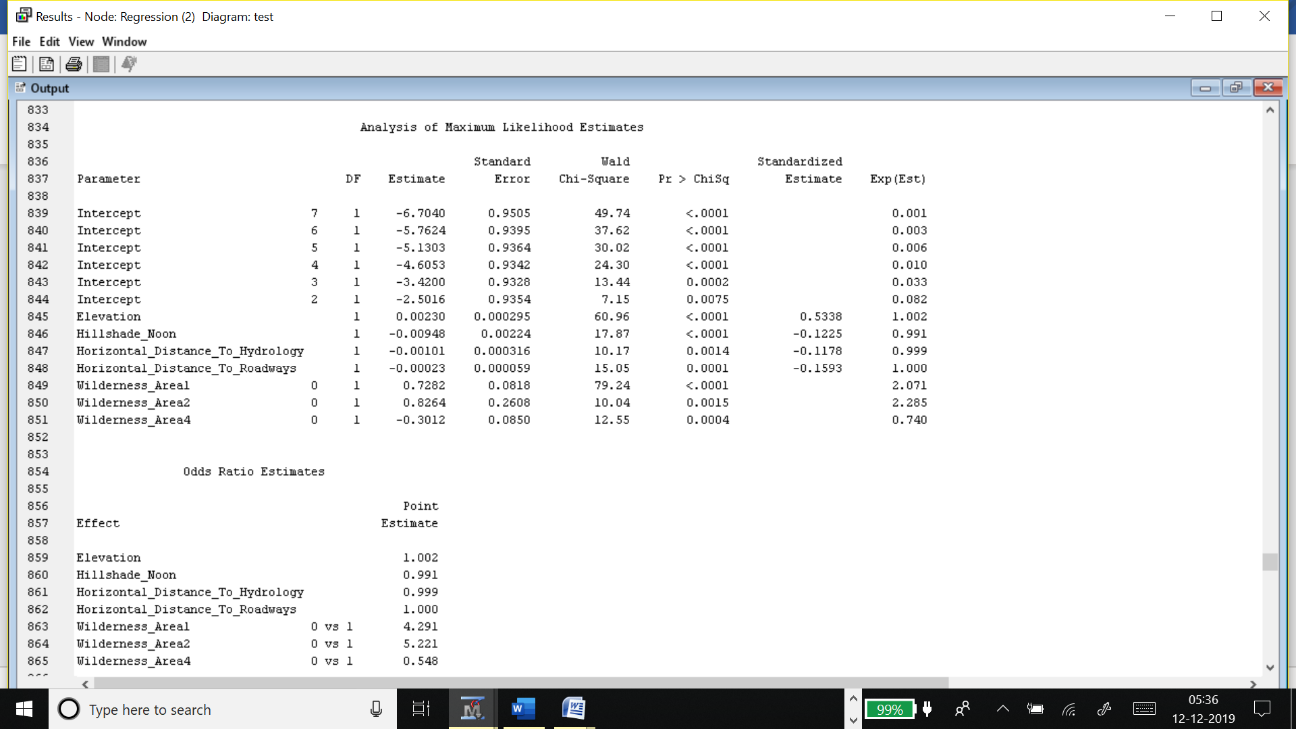


The above model is significant with a chi-square < 0.005, but there are non significant predictors which has p-value < 0.005 such as Aspect, Hillshade\_3pm, Hillshade\_9pm,slope, soil\_type1,soil\_type2,soil\_type3,soil\_type4,soil\_type5, Vertical distnace to hydrology.

We reject the value that is not significant and continue the regression until we reach to a model having all significant predictors.



The significant regression model



The above model is significant with overall p<0.05(significance level) and the significant variables such as Elevation, Hillshade\_Noon, Horizonatal\_Distance to roadways, wilderness areas 1 to 4.

Regression equation :

From the above regression, it is clear that for a one-unit increase in elevation the chance of forest cover increases exponentially by 0.00230 and for a one-unit increase in distance to roadways, the probability decreases exponentially by 0.00023, which is practical because forest grows far from the human habitats and the roadways.

Ex:

For data Intercept7=-6.7040, elevation=3298, Hillshade\_noon=222, Hor\_Distance\_to\_hydrology=134, Horizontal\_distance\_to\_roadways=4039, wilderness\_area1=1, wilderness\_Area2=0, wilderness\_area4=0.

log(p ≤ i ) = -6.7040+ (0.00230)\* (3298)– (0.00948)\*(222) – (0.00101)\*(134)-(0.00023)\*(4039)+(0.7282)\*(1)+(0.8264)\*(0)-(0.3012)\*(0) =-1.55927

P7 = e(-1.55927)=0.21029

This implies that these predictors have 20% chance of growing the aspen forest cover.

1. We can imply from the regression output that forest cover type, definetly depends on the elevation, Horizontal\_Distance\_To\_hydrology,Hillshade\_noon, HorizontalDisatnceToROadways, wilderness\_area1, wilderness\_area2, wilderness\_area3 as P-value obtained < 0.005.
2. This proves the hypothesis, that the type of forest cover type will depend on the variables above mentioned. Therefore, the type of forest cover will depend on the elevation, Horizontal\_Distance\_To\_hydrology, Hillshade\_noon, HorizontalDisatnceToRoadways, wilderness\_area1, wilderness\_area2, wilderness\_area3 as P-value obtained is less than 0.05, this practically implies that for growing a forest cover we need a water source, sufficient light, and they should be in the wilderness areas and the roadways which are far from the human habitats.

Hypothesis2: Spruce/fir forests grow in the elevated regions of more than1700 meters.

H0: Mean elevation for spruce/fir is µ1 ≤ 1700 meters.

H1: Mean elevation for spruce/fir is µ2 ≥ 1700 meters.

Mean of elevation for spruce/fir forest type is : 3140.47

Hypothesis mean = 1700 meters

Standard deviation for spruce/fir forest type is : 177.203

Sample size(n) =149

T-test:

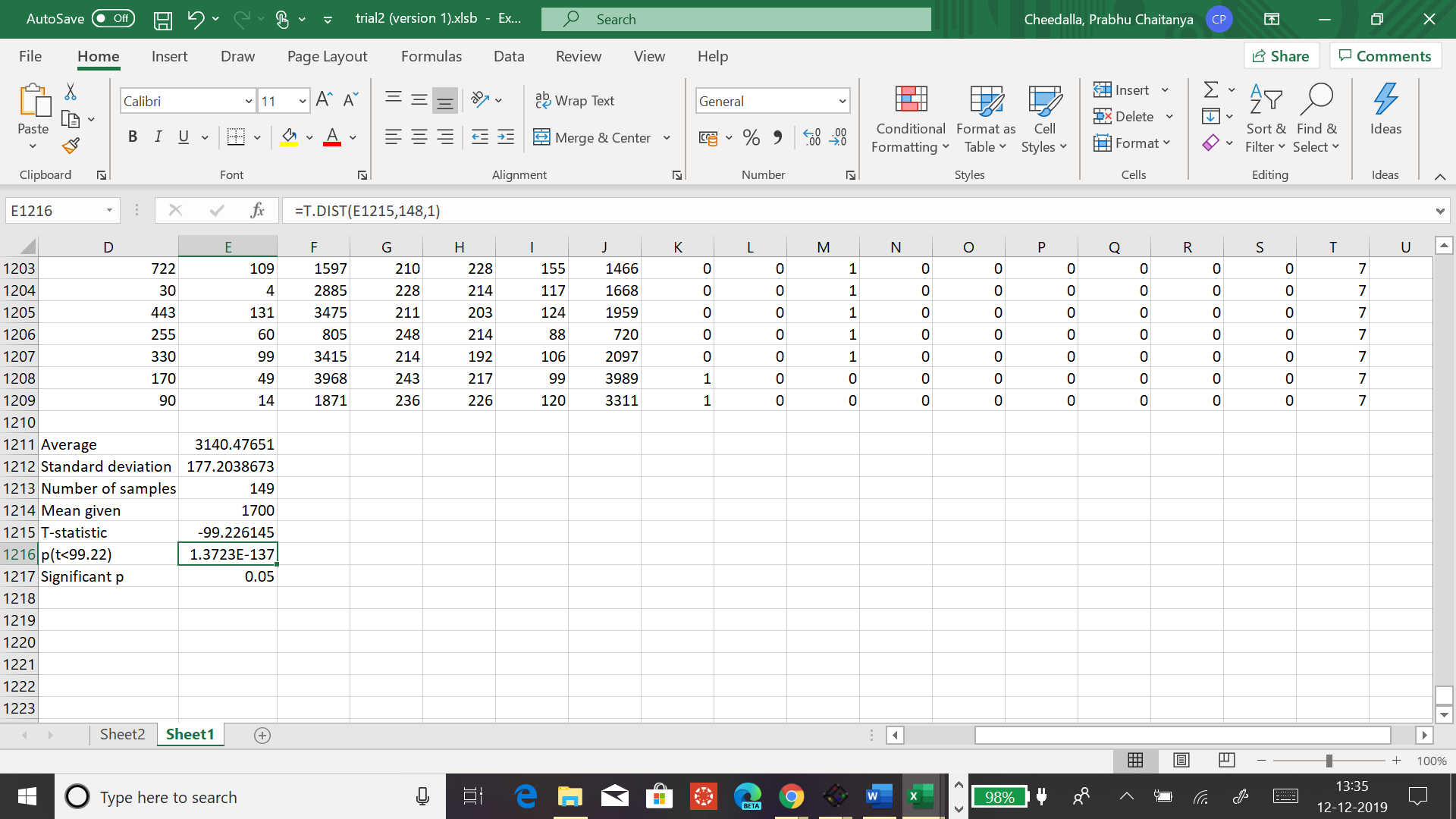
t-stat= (X̅ - µ )/(σ /√ n)

t-stat= -99.22

P(t<-99.22) with df=148 is nearly 0.00

As P-value(0.00) is less than p-significant ( 0.05), we reject the null hypothesis that, mean should be less than 1700 meters.

Therefore, it is proved that the spruce/fir forest type is grown mainly in the regions having elevations above 1700 meters. (mean of elevation ≥ 1700 meters).



Data Limitations

This data set taken from UCI repository has the data collected in the year 1998, which is so long that the climatic and geographic conditions may have changed in the past 20 years and the data set has only the effects of the geographic conditions like elevation, light intensity, soil type but not the effects of climatic conditions(like temperature, moisture, etc..,) on the forest type.

Future references

For future references if this data is analyzed along with the climatic conditions, then the model created will be most appropriate and most useful for timber farmers and timber-related companies. If this model is implemented for the agriculture farms data set, then it will be beneficial for the farmers to know which land is suitable for them to grow a specific plantation.

# References

Aravindkr. (n.d.). *R-statistical model examples*. Retrieved from http://rpubs.com/aravindkr

Blackard, J. A. (1998, August). *Forest covertype dataset*. Retrieved from https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

Blackard, J. A. (n.d.). *forest cover dataset*. Retrieved from Kaggle: https://www.kaggle.com/c/forest-cover-type-prediction

department, S. r. (n.d.). *US forest products industry*. Retrieved from https://www.statista.com/topics/1316/forest-products-industry/

Rawat, A. (n.d.). *Ordinal logistic regression*. Retrieved from https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5