G8: Aryan Raja(aryankr2), Mai Pham(maitp2), Chandler Stewart(cs111)

G8: Forest Cover Type Prediction

I. Abstract

According to climate.gov, in 2021 climate sea levels hit a new record high of 97mm above 1993 levels and continue to accelerate every year [6]. These rising sea levels are indicative of the increasingly dire state of global warming. Among the consequences of global warming is deforestation. For this reason, we have chosen to focus our project on predicting the predominant forest cover type of a forest-based solely on cartographic variables. Cartographic variables here refer to the geographic features that describe the landscape. Our dataset [8] includes observations taken from the Roosevelt National Forest of northern Colorado. We performed both 1-D and 2-D exploratory analysis to understand the nature of the data and ensure its cleanliness. This provided observations useful to engineering new features that we used to create an augmented data set. We trained various supervised learning algorithms on both the original and augmented datasets and compared their accuracy, auc, precision, recall, and f1 score. Our final results show that ensemble methods such as random forests, extreme gradient boosting, and light gradient-boosting machines are ideal for this classification task. These methods prevail over others in all previously mentioned metrics as well as training time.

II. Introduction

The project is a past Kaggle competition called Forest Cover Type which focuses on multi-classification. As input, we have features that describe the landscape and output is the classification of forest cover type (7 classes). The area of study is four wilderness areas located in the Roosevelt National Forest of northern Colorado.

III. Motivation

Successful prediction for the predominant forest cover type within a given forest subsection based solely on cartographic variables has implications for other environmental inferences. This could prove useful in predicting natural growth patterns in recovering deforested areas. We also suspect our developed technique could be used for similar predictions such as the location of natural resources, or, more generally, areas most probable to be drastically affected in the future by climate change.

IV. Related Work

This is the only problem that utilizes cartographic variables (variables derived from cartography such as elevation, aspect, etc) for classification. Due to this, the related works also focus on forest cover type prediction. One study found limited success utilizing a multi-class support vector machine classifier and K-Means clustering algorithms, receiving ~70% accuracy [1]. A separate study obtained 82% accuracy using Extremely Random decision trees. The paper concludes that the main challenge was the discrepancy in size between the train and test sets [2]. However, most works found success utilizing a mixture of ensemble techniques and feature engineering [2,3,4,5].

V. Methodology(Part 1) - Data Analysis/Cleaning

The provided data set includes 56 feature (54 training features) columns with 15120 data points for training and 565892 data points for testing. Our process went as follows:

- Step 1: Exploratory data analysis. We construct frequency tables and bar charts for categorical data and calculate percentiles, kurtosis, skewness, and histograms for numerical data
- Step 2: Cleaning data by removing any duplicate, resolving missing values, and removing outliers
- Step 3: Correlation between every two columns by using methods like Chi-square independence test, contingency tables, Pearson correlation, other hypothesis testing with visualizations
- Step 4: Feature engineering. Past works for the same project have shown significant accuracy by engineering new features [2,3]. Based on the strong disparity of numerical features, we generate a few more features: Euclidean distance to hydrology, sum/difference of elevation and vertical distance to hydrology, mean of all hillshade, sum/diff of horizontal distance to

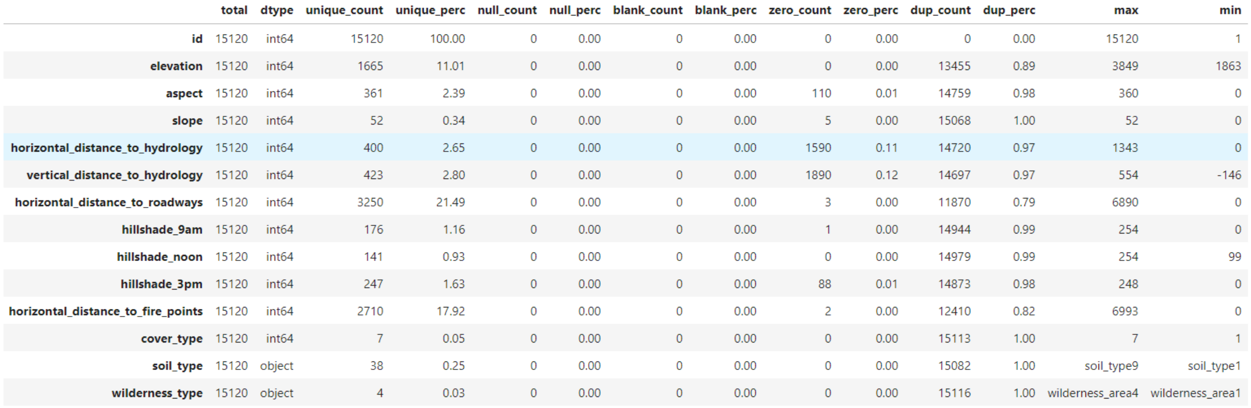
firepoints/roadways and hydrology, Euclidean distance of horizontal and vertical distance to firepoint/roadways/ hydrology, difference between hillshades. etc..

VI. Methodology(Part 2) - Model Training/Testing

- Step 1: Due to access to labeled data, we explored various supervised-learning models for multi-classification. Most past works showed best success using ensemble techniques, so that was our primary focus, however we also included several non-ensemble methods [2,3,4,5]. The techniques we utilized include: logistic regression, k-nearest neighbor, support vector machines with linear and RBF kernels, random forests , extreme gradient boosting , adaptive boosting, light gradient-boosting machine, and a deep neural network. We trained and evaluated all models ten times across both the original and augmented datasets and took the mean across all iterations of the accuracy, auc, precision, and recall, and f1 score.
- Step 2: Unsupervised Learning. We trained an AutoEncoder to visualize a reduced representation of the dataset and develop an intuition regarding the viability of unsupervised learning clustering algorithms.

VII. Empirical Results(Part 1) - Data Overview

The dataset was clean and consisted of 15120 data points after preprocessing and label encoding of the categorical variables. The dataset had 12 features, 10 numerical and 2 categorical, and 7 distinct labels that were identically distributed, with 2160 training samples for each label. **Figure 1** shows our final data overview. The first feature to note is the minimum of the vertical_distance_to_hydrology feature, which is -146. This indicates that the elevation of the tree is lower than that of the nearest surface water. Lastly, the categorical variable soil_type lacks any data points within the training set corresponding to types 7 and 15.



	total	dtype	unique_count	unique_perc	null_count	null_perc	blank_count	blank_perc	zero_count	zero_perc	dup_count	dup_perc	max	min
id	15120	int64	15120	100.00	0	0.00	0	0.00	0	0.00	0	0.00	15120	1
elevation	15120	int64	1665	11.01	0	0.00	0	0.00	0	0.00	13455	0.89	3849	1863
aspect	15120	int64	361	2.39	0	0.00	0	0.00	110	0.01	14759	0.98	360	0
slope	15120	int64	52	0.34	0	0.00	0	0.00	5	0.00	15068	1.00	52	0
horizontal_distance_to_hydrology	15120	int64	400	2.65	0	0.00	0	0.00	1590	0.11	14720	0.97	1343	0
vertical_distance_to_hydrology	15120	int64	423	2.80	0	0.00	0	0.00	1890	0.12	14697	0.97	554	-146
horizontal_distance_to_roadways	15120	int64	3250	21.49	0	0.00	0	0.00	3	0.00	11870	0.79	6890	0
hillshade_9am	15120	int64	176	1.16	0	0.00	0	0.00	1	0.00	14944	0.99	254	0
hillshade_noon	15120	int64	141	0.93	0	0.00	0	0.00	0	0.00	14979	0.99	254	99
hillshade_3pm	15120	int64	247	1.63	0	0.00	0	0.00	88	0.01	14873	0.98	248	0
horizontal_distance_to_fire_points	15120	int64	2710	17.92	0	0.00	0	0.00	2	0.00	12410	0.82	6993	0
cover_type	15120	int64	7	0.05	0	0.00	0	0.00	0	0.00	15113	1.00	7	1
soil_type	15120	object	38	0.25	0	0.00	0	0.00	0	0.00	15082	1.00	soil_type9	soil_type1
wilderness_type	15120	object	4	0.03	0	0.00	0	0.00	0	0.00	15116	1.00	wilderness_area4	wilderness_area1

Figure 1: Feature overview of training dataset

VIII. Empirical Results(Part 2) - 1D and 2D Exploratory Data Analysis

For our 1-dimensional data analysis we obtained the min, max, mean, standard deviation, and variance of all numerical features. Elevation and all features depicting horizontal distance were found to have a very high variance. Additionally, taking the n^{th} quartile across all features in increments of 0.05 showed that Vertical_Distance_To_Hydrology has a Q0.99 value of 247, however, the range of the feature: [-146,554] depicts large outliers. **Figure 2** shows the distribution of all features, most are normally distributed with either a left or right skew with the exception of Aspect where the set lacks data points corresponding to the aspect range: [150,300].Our 2-dimensional exploratory data analysis explores the correlations between features and how each feature correlates to the label. **Figure 4** depicts masked correlation heatmaps, using the magnitude of the Pearson correlation coefficient for numerical data and Cramér's V for categorical.

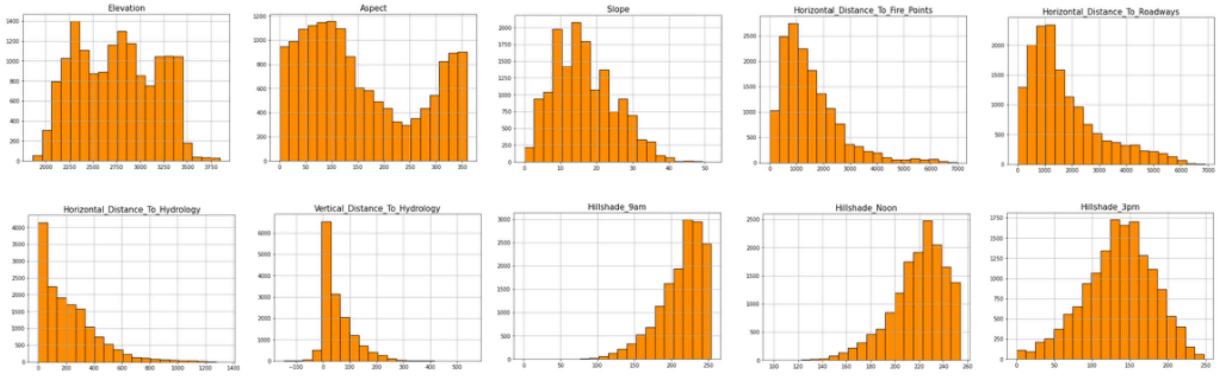


Figure 2: Histogram distribution of numerical features

From a 2-dimensional analysis between cover type and wilderness areas, we found that some cover types are specific to some wilderness area types as well as some soil types. Additionally, some numerical features illustrate vastly disparate distributions based upon cover type. These characteristics will facilitate the feature engineering process. For example, **Figure 3** shows the distribution of Elevation for each cover type.

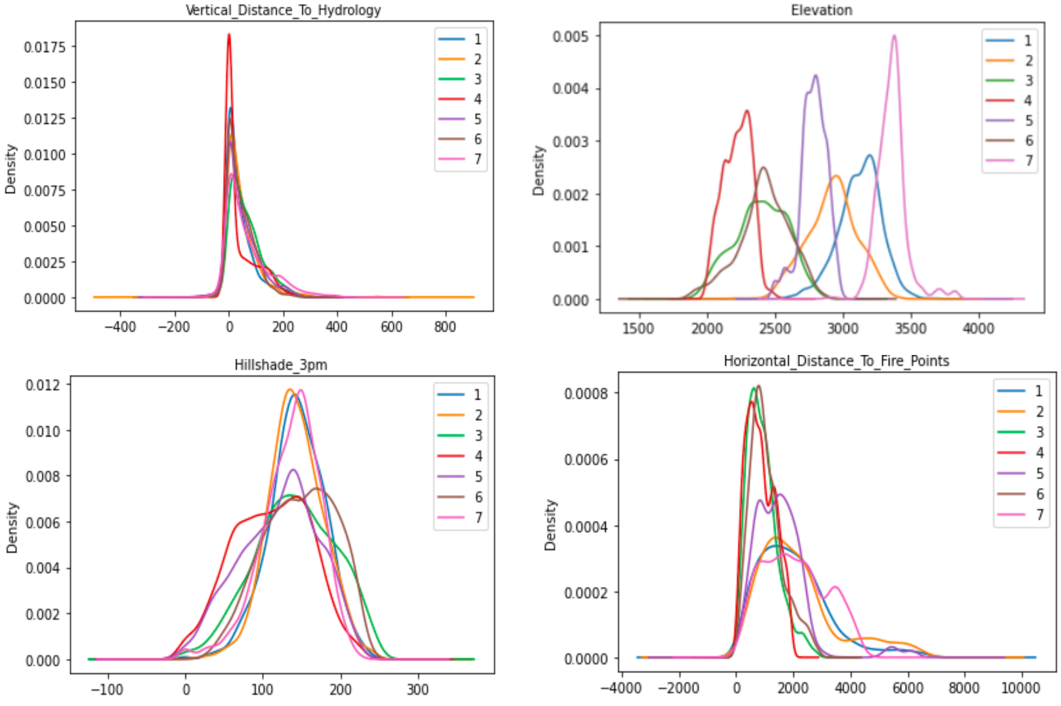


Figure 3: Distribution of four different features based on cover type. Notice the difference in the distribution of Elevation compared to others

From these 2D EDA between columns and labels, it can be seen that there are some classes significantly distinguished from each other. For example, we obviously see that each class is quite separated. This would be a strong feature and we make use of this feature to generate more features related to it. Another key point is that in vertical distance to hydro, the density of class 4 is extremely high and higher than any other classes at a point near 0. Similar to other graphs, class 4 is quite distinctive from others.

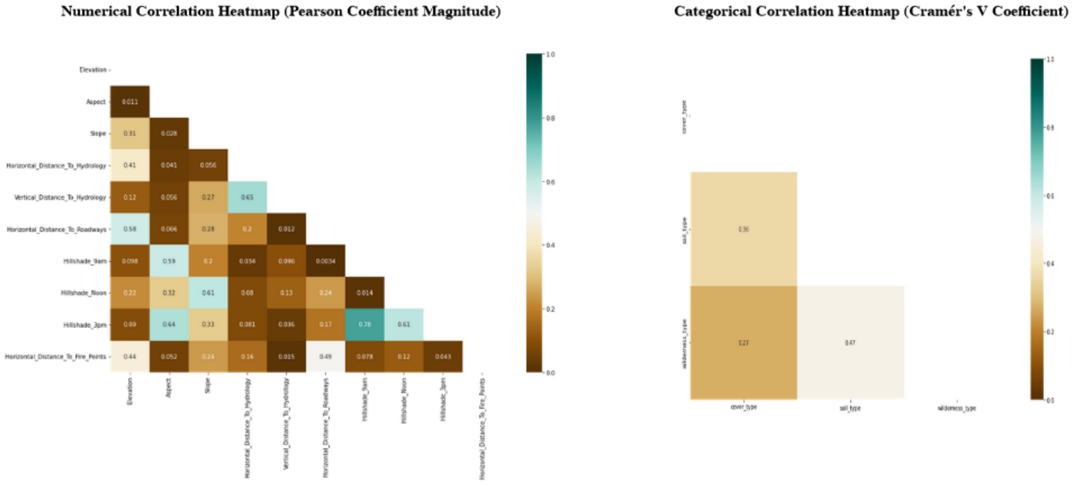
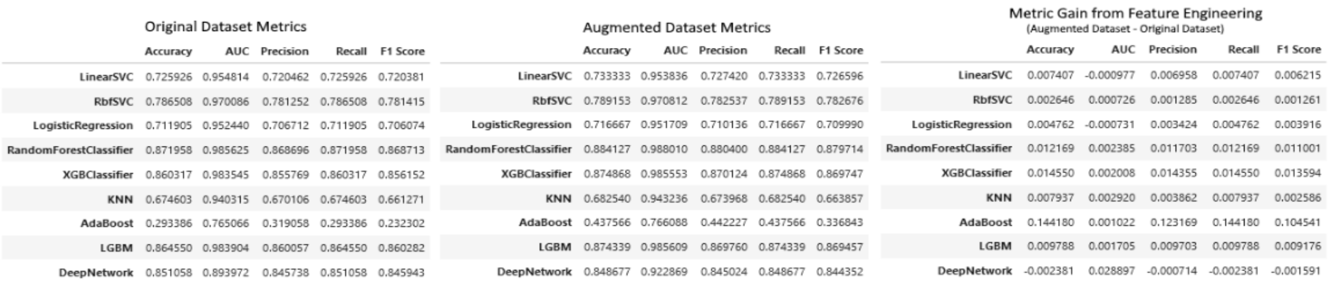


Figure 4: Correlation magnitude of numerical vs numerical (left) and categorical vs categorical (right)

IX. Empirical Results(Part 3) - Supervised Classification

Consistent with past related works, our results show ensemble techniques performing best in all metrics with the deep neural network slightly behind. **Figure 6** provides the metrics across all classifiers when trained on the original and augmented data sets as well as the difference between the two. It can be seen from the charts, three models: Extreme Gradient Boosting, Light Gradient-Boosting Machine, and Random Forests outperformed other models. Although DeepNet has achieved quite good results, this model costs a lot of time to train. Therefore, we took these top three and explored the feature importance for these models.



Original Dataset Metrics

	Accuracy	AUC	Precision	Recall	F1 Score
LinearSVC	0.725926	0.954814	0.720462	0.725926	0.720381
RbfSVC	0.786508	0.970086	0.781252	0.786508	0.781415
LogisticRegression	0.711905	0.952440	0.706712	0.711905	0.706074
RandomForestClassifier	0.871958	0.985625	0.868696	0.871958	0.868713
XGBClassifier	0.860317	0.983545	0.855769	0.860317	0.856152
KNN	0.674603	0.940315	0.670106	0.674603	0.661271
AdaBoost	0.293386	0.765066	0.319058	0.293386	0.232302
LGBM	0.864550	0.983904	0.860057	0.864550	0.860282
DeepNetwork	0.851058	0.893972	0.845738	0.851058	0.845943

Augmented Dataset Metrics

	Accuracy	AUC	Precision	Recall	F1 Score
LinearSVC	0.733333	0.953836	0.727420	0.733333	0.726596
RbfSVC	0.789153	0.970812	0.782537	0.789153	0.782676
LogisticRegression	0.716667	0.951709	0.710136	0.716667	0.709990
RandomForestClassifier	0.884127	0.988010	0.880400	0.884127	0.879714
XGBClassifier	0.874868	0.985553	0.870124	0.874868	0.869747
KNN	0.682540	0.943236	0.673968	0.682540	0.663857
AdaBoost	0.437566	0.766088	0.442227	0.437566	0.336843
LGBM	0.874339	0.985609	0.869760	0.874339	0.869457
DeepNetwork	0.848677	0.922869	0.845024	0.848677	0.844352

Metric Gain from Feature Engineering
(Augmented Dataset - Original Dataset)

	Accuracy	AUC	Precision	Recall	F1 Score
LinearSVC	0.007407	-0.000977	0.006958	0.007407	0.006215
RbfSVC	0.002646	0.000726	0.001285	0.002646	0.001261
LogisticRegression	0.004762	-0.000731	0.003424	0.004762	0.003916
RandomForestClassifier	0.012169	0.002385	0.011703	0.012169	0.011001
XGBClassifier	0.014550	0.002008	0.014355	0.014550	0.013594
KNN	0.007937	0.002920	0.003862	0.007937	0.002586
AdaBoost	0.144180	0.001022	0.123169	0.144180	0.104541
LGBM	0.009788	0.001705	0.009703	0.009788	0.009176
DeepNetwork	-0.002381	0.028897	-0.000714	-0.002381	-0.001591

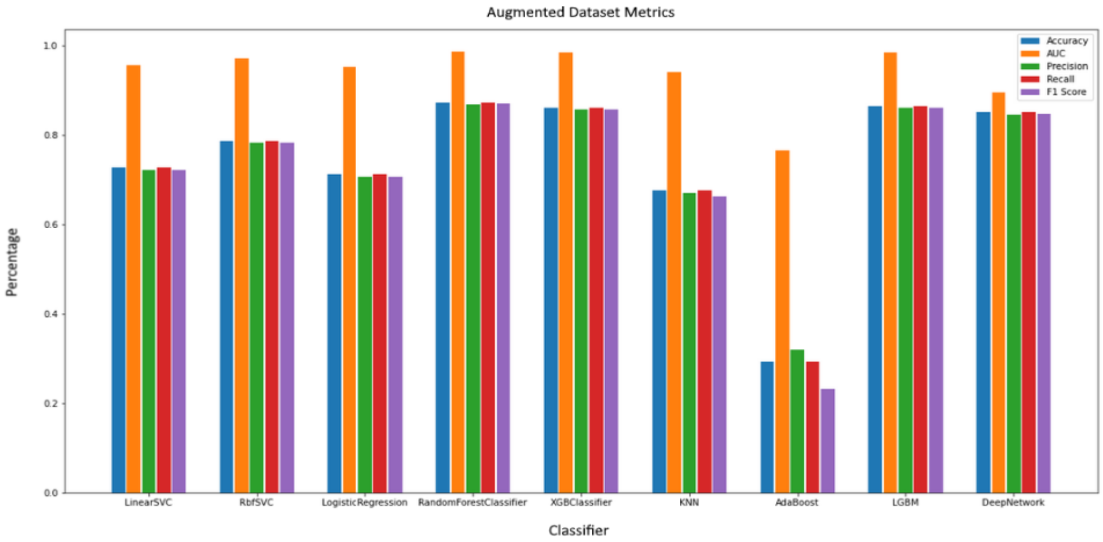


Figure 5: Accuracy, Area under the ROC Curve (AUC), Precision, Recall, and F1 score for 9 different supervised learning classification techniques. The Metric Gain from Feature Engineering table shows the effect of feature engineering on the performance of these models

Figure 6 shows the top 15 features for these models. Of note is 11 of these 15 features were engineered. With each model, we also use cross validation with StratifiedKFold to reserve the percentage of samples for each class.

	Accuracy	**AUC**	**Precision**	**Recall**
XGboost	Mean: 87.202 Min: 86.441 Max: 88.029 Std: 0.537	Mean: 98.570 Min: 98.385 Max: 98.751 Std: 0.121	Mean: 86.969 Min: 86.238 Max: 87.809 Std: 0.545	Mean: 87.202 Min: 86.441 Max: 88.029 Std: 0.537
RandomForest	Mean: 88.155 Min: 87.5 Max: 88.756 Std: 0.489	Mean: 98.861 Min: 98.770 Max: 98.969 Std: 0.077	Mean: 87.972 Min: 87.371 Max: 88.590 Std: 0.494	Mean: 88.155 Min: 87.5 Max: 88.756 Std: 0.489
LGBM	Mean: 86.462 Min: 87.765 Max: 88.756 Std: 0.265	Mean: 98.4 Min: 98.279 Max: 98.498 Std: 0.086	Mean: 85.564 Min: 85.232 Max: 86.122 Std: 0.277	Mean: 86.462 Min: 87.765 Max: 88.756 Std: 0.265

Table 1: Shows the standard deviation and statistics of the training process using cross validation with Stratified K-Fold (k=5).

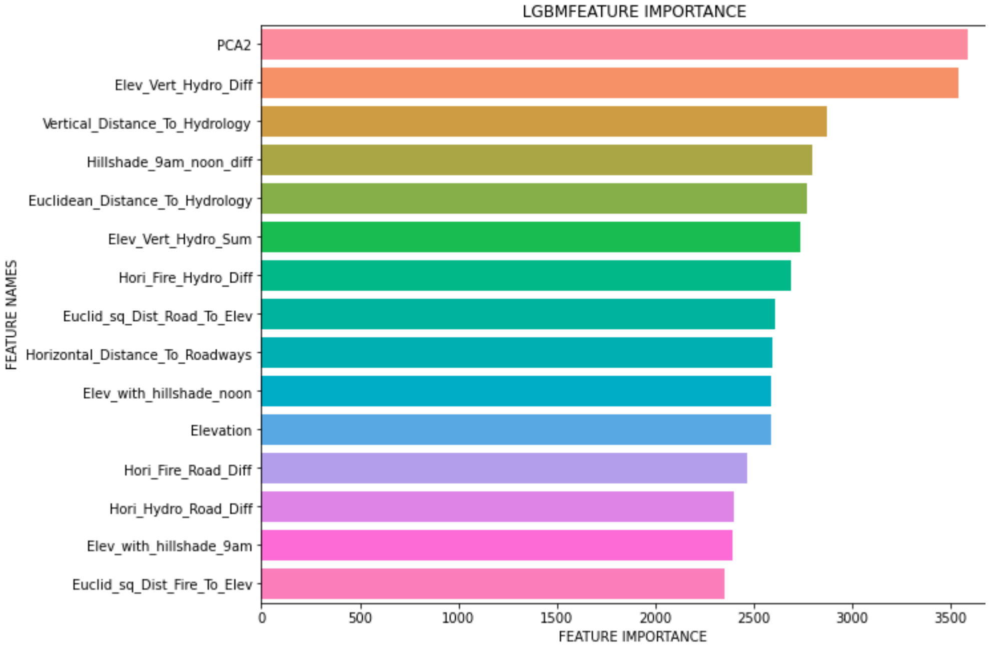


Figure 6: Features ordered on LGBM Feature Importance (greatest to lowest importance)

We also measure the multi class log loss, shown in **Figure 7**

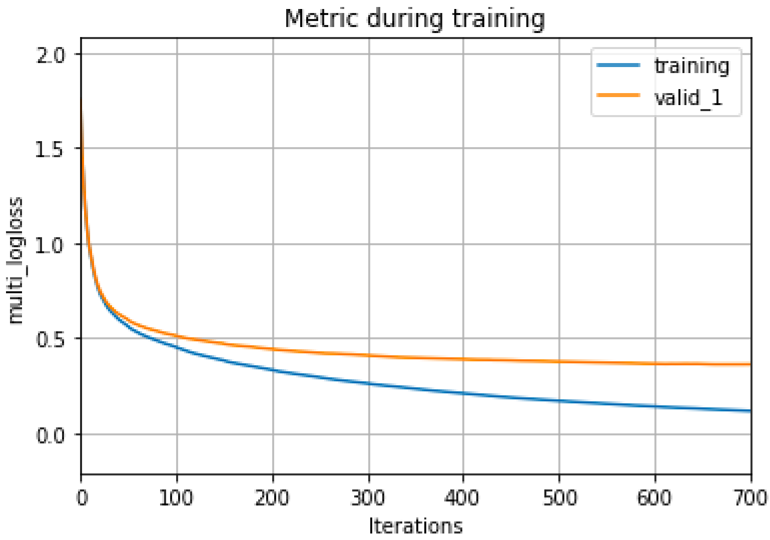


Figure 7: Training and Validation Loss Graph

X. Empirical Results(Part 4) - Unsupervised Learning

We trained an AutoEncoder with the augmented dataset and took the latent representation to visualize how well the feature engineering could aid clustering with algorithms such as K-Means. **Figure 8** shows how well our information is preserved when reducing our augmented dataset from a 77-dimensional space to a 3-dimensional space. There seems to be some clear separations between several of the classes even in such a low dimension space.

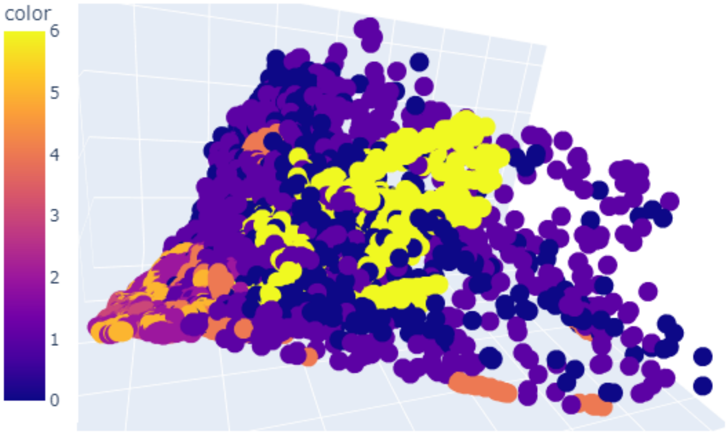


Figure 8: AutoEncoder latent representation of dataset displaying a clear separation of classes in a 3-dimensional space

XI. Conclusion/Discussions

We have shown that it is possible to predict forest cover type based purely on cartographic variables with relative accuracy. The separate feature distributions in **Figure 3** are evidence of the fact that there exists distributions that are categorically unique. We believe that forest cover type prediction is not unique in that a combination of these cartographic variables can be utilized for accurate prediction for other problems. Based on our supervised classification results as well as our unsupervised latent representation, we believe clustering techniques of other cartographic variable datasets could provide insight into other natural phenomena such as natural resource locations.

References

[1] Crain, K. (2014). *Classifying Forest Cover Type using Cartographic Features*. Cs229.stanford.edu. Retrieved 3 October 2022, from https://cs229.stanford.edu/proj2014/Kevin%20Crain,%20Graham%20Davis,%20Classifying%20Forest%20Cover%20Type%20using%20Cartographic%20Features.pdf.

[2] Idelbayev, Y. *Assignment 1. Predicting cover of forest*. Cseweb.ucsd.edu. Retrieved 3 October 2022, from https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/wi15/Yerlan_Idelbayev.pdf.

[3] Kolasa, T., & Raja, A. *Forest Cover Type Classification Study*. Rstudio-pubs-static.s3.amazonaws.com. Retrieved 3 October 2022, from https://rstudio-pubs-static.s3.amazonaws.com/160297_f7bcb8d140b74bd19b758eb328344908.html.

[4] Petrusevich, D. (2021). *Models for dominating forest cover type prediction*. Iopscience.iop.org. Retrieved 3 October 2022, from https://iopscience.iop.org/article/10.1088/1755-1315/677/5/052119/pdf.

[5] Vitebskyy, M. (2019). *Using Machine Learning to Determine a Forest Cover Type*. Medium. Retrieved 3 October 2022, from https://medium.com/cuny-csi-mth513/using-machine-learning-to-determine-a-forest-cover-type-9456eb60635e.

[6] Lindsey, R. (2022). *Climate Change: Global Sea Level*. climate.gov. Retrieved 3 October 2022, from https://www.climate.gov/news-features/understanding-climate/climate-change-global-sea-level.

[7] *How hillshade works*. How HillShade works-ArcGIS Pro | Documentation. (n.d.). Retrieved November 2, 2022, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/3d-analyst/how-hillshade-works.htm

[8] J. Blackard, Covertype. UCI Machine Learning Repository, 1998.