Mining New York City Airbnb Listings

Husna Sayedi
UC Riverside
hsaye001@ucr.edu

Kevin Vega
UC Riverside
kvega008@ucr.edu

Shahab Geravesh
UC Riverside
sgera001@ucr.edu

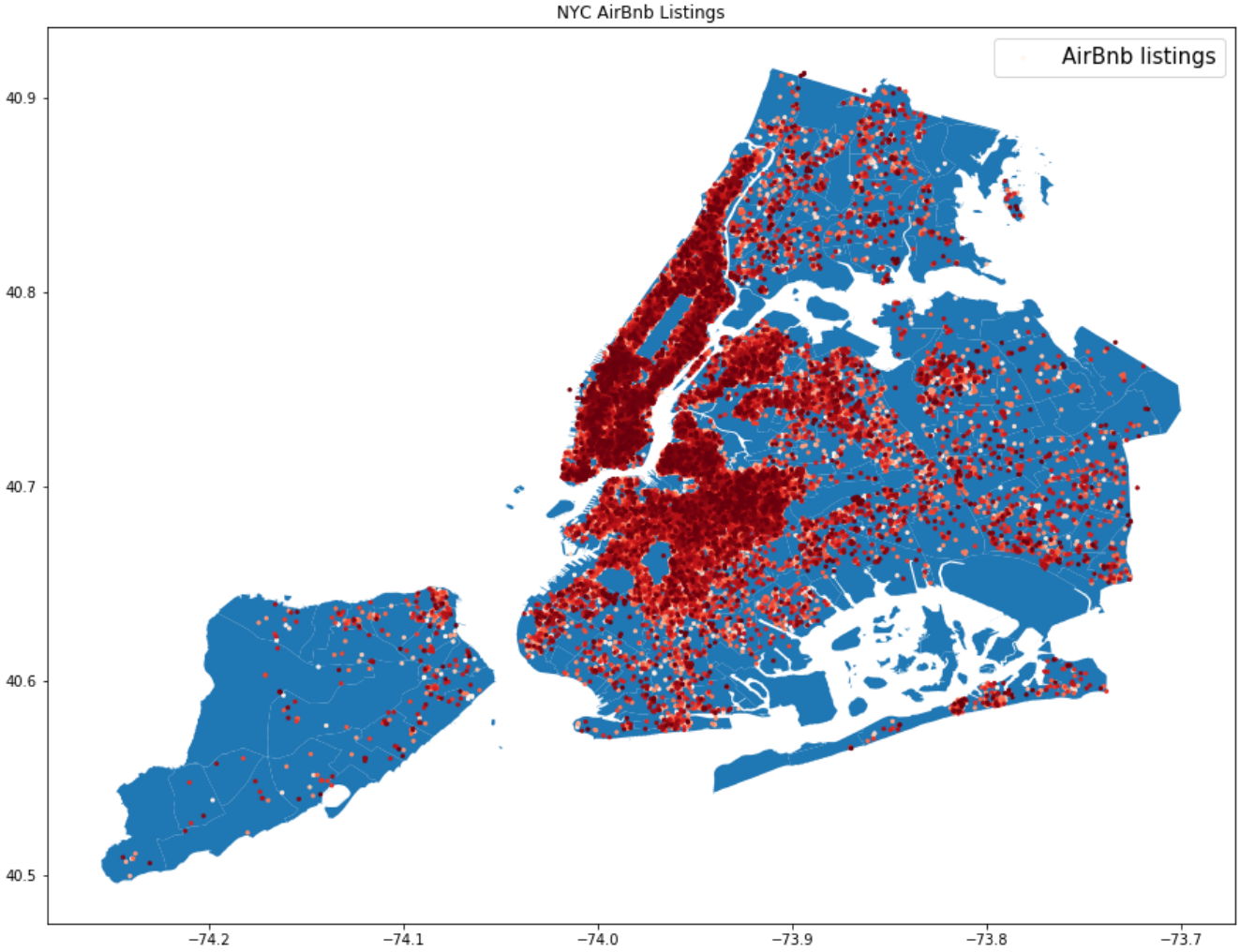

Figure 1: NYC Airbnb Listings

ABSTRACT

As we live in an increasingly globalized society, Airbnb has become one of the most popular housing or renting accommodation marketplace. The company provides a platform which connects both hosts and renters to either rent out their space or book a listing. Airbnb is offered today in over 81,000 cities and 191 countries worldwide [0]. This immense presence leads to questions of how blah blah. In this project, we wish to examine the listing of Airbnb listings in New York City to extract knowledge and provide forecasting. We use a series of data mining and machine learning methodologies to explore the dataset to discover insights and make listing predictions.

KEYWORDS

datasets, neural networks, clustering, data mining, k-means, Dimensionality Reduction Factor Analysis

1 INTRODUCTION

With the ever-increasing reliance and abundance of big data, data mining has become one of the most widely used methods to extract

information from large datasets. Often, data mining is referred to as knowledge discovery of data (KDD). This consists of the extraction of non-trivial and previously unknown patterns of knowledge from large amounts of data. The KDD pipeline includes the processing of raw data, including data integration, cleaning, feature selection, and dimensionality reduction. Following this is the analysis, which often includes pattern discovery (and association and correlation), classification, clustering, and outlier analysis. Post-processing concludes the pipeline, consisting of evaluation, interpretation, and visualization.

There are various techniques of data mining. Choosing a technique depends on the problem one is trying to solve. The methods we have chosen to implement include K-Means Clustering, Factor analysis, and Recurrent Neural Networks, gradient descent, and principal component analysis. Using these methods, we intend to propose practical recommendations for Airbnb prospective tenants and forecast the popularity or desirability of certain neighbourhoods, room types, listings, etc.

2 RELATED WORK

Supervised and unsupervised machine learning and data mining methods have been around for quite some time. Existing works pertaining to unsupervised machine learning and data mining techniques provide background knowledge and useful methods relating to our problem. Specifically, we examined related studies and research regarding neural networks, k-means clustering, prediction methodologies, and forecasting using support vector machines.

Different flavors of neural networks have been used to predict housing prices in the real estate market. In his paper, author Itedal Sabri Hashin Bahia uses the algorithms Feed-Forward Backpropagation (FFBP) network and a Cascade Forward Back-propagation (CFBP) network [2]. CFBP is similar to FFBP network in the fact that both are using back-propagation for updating the weights. The main difference between the two algorithms is that in CFBP network, each layer neurons relates to all previous layer neurons, thus resulting in more connections than a FFBP.

Bahia's methodology process consisted of feature selection, preprocessing, transformation, data mining, and interpretation/evaluation. After running the models, CFBP gave the best results since the MSE was less than that of FFBP. The pros of CFBP over FFBP is not only the reduced MSE rates, but also the number of epochs used was significantly less than that of FFBP, giving an overall more efficient algorithm. For our case, we can compare the FFBP method directly to predict raw prices of new listings or forecast other interesting attributes [1].

In a second paper, authors Huda Hamdan Ali and Lubna Emad Kadhum explore clustering in data mining and image processing [1]. The authors review the pros and cons of the algorithm and its implementation steps. They describe that there are many benefits of using k-means clustering, namely its efficiency in discovering underlying patterns in unsupervised learning. The algorithm has many applications and real-world implementations, its fast, and simple. We will go further into the algorithmic detail in the below sections.

Software Project Management is a tool to measure the progress and satisfaction of a software project. To predict the software time, it is essential to have a model that can predict accurately. Nowadays, leveraging Machine learning techniques can help scientists to build a model to predict as accurately as possible. Gaussian Process achieves higher total performance compared to other algorithms. Besides, leveraging M5P is beneficial for the prediction of a numeric target variable.

Recently, modern companies build workflows to observe the performance of their workers. Having a planned workflow is useful to accomplish the company's goal in a predictable way. Manually extorting workflows from a data-set is a tedious task. Therefore, Scientists strive to automate the process of workflow extraction by using natural language data-sets. An example of automating the workflow is Apache Software Foundation email archives, which enable the partners to connect and contribute from different parts of the world.

Predicting rental prices for Airbnb In New York is challenging due to the myriad factors that influence the demand for such properties. Determining the important factors can help better develop models to make better predictions. The paper, Housing price forecasting based on genetic algorithm and support vector machine, by Gu et al. explores a technique for determining the optimal parameters for an SVM using a genetic algorithm. An initial population composed of 20 chromosomes is gender is generated randomly. Each chromosome consists of 3 parameters; the kernel function parameter Sigma, the penalty parameter C, and the insensitive loss parameter epsilon. The model is then trained with each chromosome and the fitness of each chromosome is calculated. Those chromosomes which pass the fitness test pass their genes along to a new generation. This process continues until they stop criteria is met and the optimal parameters are found. The study found that genetic algorithms determine the SVM optimization parameters in less time than the most commonly used method, Grid algorithm, and performed better when forecasting prices.

Utilizing a recurrent neural network is appropriate for data where the connections between nodes form directed graph along a temporal sequence. The Airbnb data set consists of rental data spanning almost a decade. The study, Recurrent neural network for forecasting next 10 year loads of 9 Japanese utilities, by Bahman Kermanshahi, demonstrated good potential for recurrent neural networks have forecasting on datasets with many inputs. Kermanshahi's approach makes use of a contribution factor to determine levels of influence of selected inputs on the output. Twelve economic factors were used as inputs, as well as load and weather data, the output layer is the maximum load for a given year. Context neurons are used to communicate feedback from one year to the next. In the case of our project, prices of Airbnb rentals are not the only factors that change overtime. Incorporating temporal feedback for other factors that change over time such as weather, reviews per month, and other socioeconomic factors should yield better prediction capabilities.

Each of these previous two papers demonstrate the applicability of machine learning and KDD techniques to housing data.

3 PROPOSED METHODS

3.1 K-Means Clustering

Our first implementation was a K-Means Clustering algorithm. in order to cluster similar insights such as popular neighborhoods, finding the most/least expensive neighborhoods, finding which room type is the most desirable, etc. Because we have gotten familiar with the dataset, we can choose our initial k values using either intuition or a technique known as the elbow method, i.e. the number of distinct neighborhoods, the number of distinct room types, etc.

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Figure 2: K-Means Clustering Pseudocode (https://stanford.edu/~cpiech/cs221/img/kmeansMath.png).

The K-Means algorithm takes an input parameter (k) and partitions data into a set of k clusters that minimizes intra-class similarity and minimizes inter-class similarity. Cluster similarity is measured in terms of the mean value of a given cluster, also known as the centroid. The algorithm works by initialized randomly selected k which represents the mean of each (k) clusters. A data point is assigned to the most similar cluster using a distance metric between the data point and the cluster mean. Next, a new mean is computed for each cluster. These steps are repeated until a specific criterion is met or when all data points have been classified. Figure 2 shows these steps as a pseudo-code of the K-Means Clustering Algorithm:

Often times choosing K can be done through intuition of the data or prior knowledge of groupings of data points. However, it is not always the case. In times where we have no prior intuition of the natural groupings of data points, we can use the elbow method technique to choose an optimal K value. The elbow method uses a range of given K values and computes Within-Cluster-Sum-of-Squares (WCSS). The resultant plots K values against WCSS values. The optimal K value becomes the value(s) where the plot is bent the most, similar to an elbow, hence the name. Figure 3 displays the elbow plot with K=8 a subset of the Airbnb dataset, consisting of listings from Staten Island.

From Figure 3, we can see that K values which correspond to the sharpest bend in the plot are values K = 2 and 3. Hence, the elbow method dictates the optimal K values are 2 or 3, for the given particular data subset.

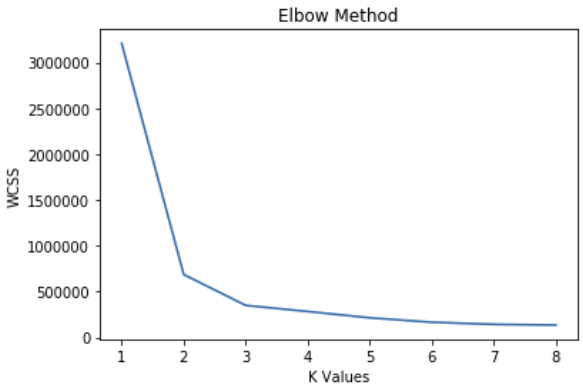

Figure 3: Elbow Method with K=8 Clusters

3.2 Neural Network

A neural network algorithm attempts to mimic biological systems to find patterns and make predictions. A neural network is a directional graph composed of nodes called neurons which communicate with each other. Neurons are organized into layers and the connections between the neurons carry information between them. Every neural network is composed of an input layer, a number of hidden layers, and an output layer. The size of input and output layers is determined by the data. Information fed through the graph is processed by the neurons which predict a result. During training the predicted value is compared to a known value. The error rate across the data set is calculated. The network minimizes this loss by penalizing the neurons that predicted an incorrect result. The connections between neurons that predict correct values are given more weight.

Much of the tuning of neural networks involves adjusting the number of hidden layers and the number of neurons per layer. The effects of these parameters on the accuracy and loss of the neural network in predicting the location of the Airbnb listings was investigated. Several models were created, with varying numbers of hidden layers and neurons per layer. Models of 1-4 layers were created, each run varying the number of neurons per layer, first 8, then 16, and finally 32 neurons per layer.

3.3 Factor Analysis

Factor analysis is a statistical method to reduce the dimensions of a high dimensional data-set by analyzing correlated variables and choosing a few most common factors to conduct data analysis. First, we drew a scatter plot and constructed a correlation matrix to measure the correlations between variables. We observed that the price is positively correlated with longitude, which means as we travel to the West, price increases. For example, Manhattan is the costliest area in New York, and it is located on the West side of the city. Latitude has a minor impact on the price, but we can see that the northern part of the city is more expensive. Price is negatively correlated with the number of reviews and the number of reviews per month. We can conclude that the reviews can lower the price. We observed that price, minimum nights, number of reviews, last

review, calculated host listing count have skewed distributions. To lower the skewness, we have transformed the data logarithmically.

Matrix.png

Correlation Matrix[a]

		logprice	logAvail3651	logcalcHostlistcount1	Logreviewspermonth1	LogNumberofreviews1	LogMinights1
Correlation	logprice	1.000	.096	.051	-.060	-.065	.063
	logAvail3651	.096	1.000	.342	.366	.292	.152
	logcalcHostlistcount1	.051	.342	1.000	-.034	-.096	.345
	Logreviewspermonth1	-.060	.366	-.034	1.000	.761	-.294
	LogNumberofreviews1	-.065	.292	-.096	.761	1.000	-.221
	LogMinights1	.063	.152	.345	-.294	-.221	1.000
Sig. (1-tailed)	logprice		.000	.000	.000	.000	.000
	logAvail3651	.000		.000	.000	.000	.000
	logcalcHostlistcount1	.000	.000		.000	.000	.000
	Logreviewspermonth1	.000	.000	.000		.000	.000
	LogNumberofreviews1	.000	.000	.000	.000		.000
	LogMinights1	.000	.000	.000	.000	.000	

Figure 4: Correlation of Normalized Variables

Statistics.png

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
logprice	2.05313341226	.303207608969	48715
logAvail3651	1.29840871062	1.069662035254	48715
logcalcHostlistcount1	.47219976340	.373349128743	48715
Logreviewspermonth1	.23514351852	.250638103482	48715
LogNumberofreviews1	.86392914264	.679224246521	48715
LogMinights1	.65171570379	.377911222291	48715

In Figure 5, Log(Availability 365 +1) has the largest standard deviation which means most of the numbers are far from the average and they are spread out. On the other hand, Log(reviews per month+1) has the least standard deviation meaning the data-points are close to the average.

plot.png

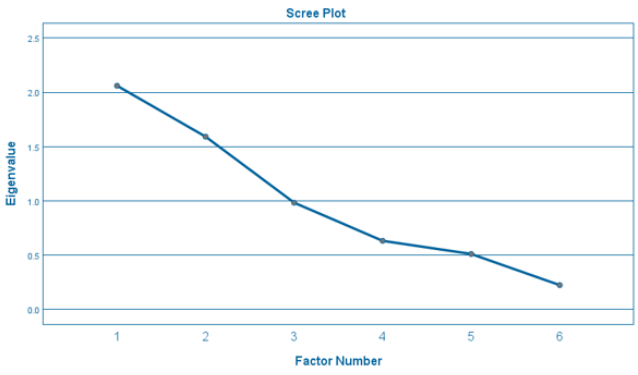

Figure 5: Scree Plot

Variance explained.png

Total Variance Explained

Factor	Initial Eigenvalues Total	% of Variance	Cumulative %	Extraction Sums of Squared Loadings Total	% of Variance	Cumulative %	Rotation Sums of Squared Loadings[a] Total
1	2.060	34.338	34.338	1.795	29.924	29.924	1.781
2	1.592	26.525	60.863	.996	16.605	46.529	1.032
3	.983	16.383	77.246				
4	.632	10.525	87.772				
5	.510	8.507	96.279				
6	.223	3.721	100.000				

Extraction Method: Maximum Likelihood.

a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

Communalities[a]

	Initial	Extraction
logprice	.021	.019
logAvail3651	.297	.434
logcalcHostlistcount1	.223	.449
Logreviewspermonth1	.631	.946
LogNumberofreviews1	.586	.614
LogMinights1	.227	.330

Extraction Method: Maximum Likelihood.

a. One or more communalitiy estimates greater than 1 were encountered during iterations. The resulting solution should be interpreted with caution.

Figure 6: Communalities

Scree plot shows eigenvalue V.S Component number. From Figure 6 and Figure 7 we can see the sharpest bend is occurring from Component 1 to component 2 or component 3. In addition, extraction sums of squared loading indicated that the extracted components explain 47 Percent of variability in the original 6 variables. Therefore, we can reduce the complexity of the data-set by using only component 1 and component 2, but we should be aware of 53 percent of loss information.

In Figure 8, communalities show the amount of variance in each variable and the extraction communalities estimates the variance in each variable. Communalities in Log(reviews per month +1) and Log(Number of reviews+1) are the highest which means these two components represent the variables well. In contrast, low communalities such as Log(Price) is not well represented in the common factor space.

4 EXPERIMENTAL EVALUATION

4.1 K-Means Clustering Results

Subsets of the Airbnb dataset were ran using the K-Means algorithm. We broke the data into its distinct neighbourhood groups consisting of Manhattan, Staten Island, Queens, and Brooklyn. From there, we wished to assess how and if price was correlated with popularity of the listings per distinct room type, which includes homes/apartments, private rooms, and shared rooms. We compared features such as 'number of reviews' and 'availability 365' against the normalized price per listing and compared the results. Figure 4 reveals the output of the K-Means algorithm comparing the normalized price versus the number of reviews of the room type home/apartment for neighborhoods (a) Queens and (b) Staten Island. As shown, the Figure 4(a) shows 7 clusters and Figure 4(b) shows 6 clusters. Figure 4(b) reveals that the cheaper listings (clusters 3 and 6) received more reviews, hence are generally more popular than the more expensive listings. The same correlation appears to be true in Figure 4(a) as well - clusters 5, 6, and 4 all have the most dense clusters.

4.2 Neural Network Results

The accuracy and loss of the neural network models increased and decreased, respectively, as the number of layers increased as well as when the number of neurons per layer increased. The largest network approached accuracy of nearly 80

5 DISCUSSION AND CONCLUSION

For the K-Means algorithm, although the results may provide some insights or visualizations, we conclude that the algorithm doesn't perform well with our dataset. Data used in this algorithm had

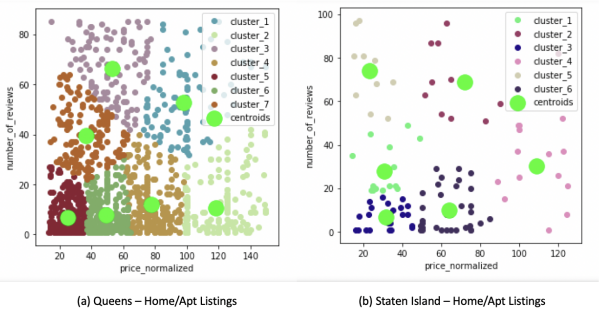

Figure 7: Listing clusters of Queens and Staten Island with Features Normalized Price and Number of Reviews

Group-4-hidden-32-neurons-acc.png

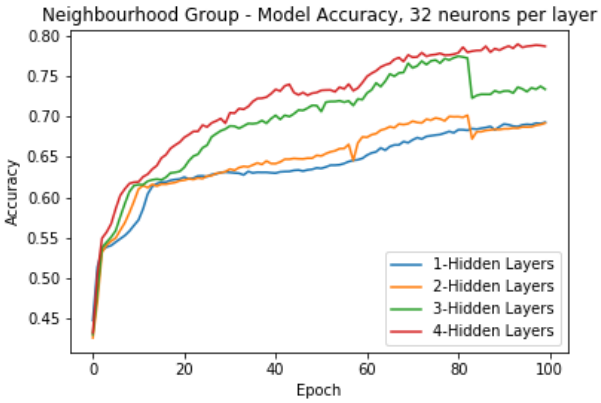

Figure 8: Accuracy of Neural Network with 4-hidden layers with 32 neurons per layer

to be cleaned thoroughly using statistical methods, as K-Means is sensitive to noise and outliers. Furthermore, The clusters do not appear to have high intra-class similarities and low inter-class similarities. Although some subsets of data showed some decent clusters, overall the results are not noteworthy. A few ways to to improve this algorithm includes analyzing more than 2 features at a time or implementing K++ means [source] to initiate better clusters initially.

For the neural network, bigger networks tended to perform better. Some improvements may included the use of GPU acceleration to speed up the computation, as large number of epochs took quite a while to compute.

[7] [1] [2] [3] [4] [5] [7] [6]

REFERENCES

[1] Huda Hamdan Ali and Lubna Emad Kadhum. [n.d.]. K- Means Clustering Algorithm Applications in Data Mining. *Semantic Scholar* ([n. d.]). https://pdfs.semanticscholar.org/a430/da239982e691638b7193ac1947da8d0d241b.pdf

[2] Itedal Sabri Hashim Bahia. 2013. A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. *International Journal of Intelligence Science* 03 (April 2013), 162–169. https://doi.org/10.4236/ijis.2013.34017

[3] Mingcang Zhu Jirong Gu and Liuguangyan Jiang. 2011. Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications* 38, 4 (2011), 3383–3386. https://doi.org/10.1016/j.eswa.2010.08.123

[4] Bahman Kermanshahi. 1998. Recurrent neural network for forecasting next 10 years loads of nine Japanese utilities. *Neurocomputing* 230, 1-3 (1998), 125–133. https://doi.org/10.1016/s0925-2312(98)00073-3

[5] Chris Piech and Andrew Ng. [n.d.]. *K Means.* Retrieved December 9, 2019 from https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

[6] Tirthajyoti Sarkar. 2019. *Clustering metrics better than the elbow-method.* Retrieved December 9, 2019 from https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6

[7] Tian Bo Lu Wan Jiang Han, Li Xin Jiang and Xiao Yan Zhang. 2015. Comparison of Machine Learning Algorithms for Software Project Time Prediction. *International Journal of Multimedia and Ubiquitous Engineering* 10, 9 (2015), 1–8. https://doi.org/10.14257/ijmue.2015.10.9.01

Group-4-hidden-32-neurons-loss.png

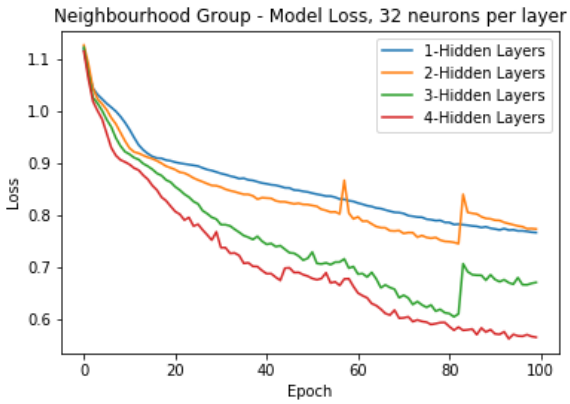

Figure 9: Loss of Neural Network with 4-hidden layers with 32 neurons per layer