Machine Learning Project Status Report

Shalini Hemachandran sxh163230

Bhakti Khatri brk160030

Sabarish Nadarajan sxn164530

Somya Singh sxs161331

Partha De pxd141430

Introduction

A regression problem involving the **Prediction of House Prices** (advanced version of Boston Housing Data) has been chosen for the project. Our aim is to make it to the leaderboards in Kaggle. We have divided the project into three parts, namely

1. Feature Engineering
2. ANN Training which involves parameter tuning, validation and evaluation
3. Evaluation and Comparison with established results and tuning the model for best results

Dataset Details

Data Set Characteristics	Multivariate
Attribute Characteristics	33 Numeric and 46 Categorical
Associated Tasks	Regression
Number of Instances	1460
Number of Attributes	80
Missing Values	6965

Since the dataset contains categorical attributes, they must be converted into numerical attributes before they can be used to train ANN. Also, the attributes must be scaled.

Below is a list of the eighty input attributes.

Attributes	**Type**	**Description**
MSSubClass	Categorical	Identifies the type of dwelling involved in the sale.
MSZoning	Categorical	Identifies the general zoning classification of the sale.
LotFrontage	Numeric	Linear feet of street connected to property
LotArea	Numeric	Lot size in square feet
Street	Categorical	Type of road access to property

Alley	Categorical	Type of alley access to property
LotShape	Categorical	General shape of property
LandContour	Categorical	Flatness of the property
Utilities	Categorical	Type of utilities available
LotConfig	Categorical	Lot configuration
LandSlope	Categorical	Slope of property
Neighborhood	Categorical	Physical locations within Ames city limits
Condition1	Categorical	Proximity to various conditions
Condition2	Categorical	Proximity to various conditions (if more than one is present)
BldgType	Categorical	Type of dwelling
HouseStyle	Categorical	Style of dwelling
OverallQual	Categorical	Rates the overall material and finish of the house
OverallCond	Categorical	Rates the overall condition of the house
YearBuilt	Numeric	Original construction date
YearRemodAdd	Numeric	Remodel date (same as construction date if no remodeling or additions)
RoofStyle	Categorical	Type of roof
RoofMatl	Categorical	Roof material
Exterior1st	Categorical	Exterior covering on house
Exterior2nd	Categorical	Exterior covering on house (if more than one material)
MasVnrType	Categorical	Masonry veneer type
MasVnrArea	Numeric	Masonry veneer area in square feet
ExterQual	Categorical	Evaluates the quality of the material on the exterior
ExterCond	Categorical	Evaluates the present condition of the material on the exterior
Foundation	Categorical	Type of foundation
BsmtQual	Categorical	Evaluates the height of the basement
BsmtCond	Categorical	Evaluates the general condition of the basement

BsmtExposure	Categorical	Refers to walkout or garden level walls
BsmtFinType1	Categorical	Rating of basement finished area
BsmtFinSF1	Numeric	Type 1 finished square feet
BsmtFinType2	Categorical	Rating of basement finished area (if multiple types)
BsmtFinSF2	Numeric	Type 2 finished square feet
BsmtUnfSF	Numeric	Unfinished square feet of basement area
TotalBsmtSF	Numeric	Total square feet of basement area
Heating	Categorical	Type of heating
HeatingQC	Categorical	Heating quality and condition
CentralAir	Categorical	Central air conditioning
Electrical	Categorical	Electrical system
1stFlrSF	Numeric	First Floor square feet
2ndFlrSF	Numeric	Second floor square feet
LowQualFinSF	Numeric	Low quality finished square feet (all floors)
GrLivArea	Numeric	Above grade (ground) living area square feet
BsmtFullBath	Numeric	Basement full bathrooms
BsmtHalfBath	Numeric	Basement half bathrooms
FullBath	Numeric	Full bathrooms above grade
HalfBath	Numeric	Half baths above grade
Bedroom	Numeric	Bedrooms above grade (does NOT include basement bedrooms)
Kitchen	Numeric	Kitchens above grade
KitchenQual	Categorical	Kitchen quality
TotRmsAbvGrd	Numeric	Total rooms above grade (does not include bathrooms)
Functional	Categorical	Home functionality (Assume typical unless deductions are warranted)
Fireplaces	Numeric	Number of fireplaces
FireplaceQu	Categorical	Fireplace quality
GarageType	Categorical	Garage location

GarageYrBlt	Numeric	Year garage was built
GarageFinish	Categorical	Interior finish of the garage
GarageCars	Numeric	Size of garage in car capacity
GarageArea	Numeric	Size of garage in square feet
GarageQual	Categorical	Garage quality
GarageCond	Categorical	Garage condition
PavedDrive	Categorical	Paved driveway
WoodDeckSF	Numeric	Wood deck area in square feet
OpenPorchSF	Numeric	Open porch area in square feet
EnclosedPorch	Numeric	Enclosed porch area in square feet
3SsnPorch	Numeric	Three season porch area in square feet
ScreenPorch	Numeric	Screen porch area in square feet
PoolArea	Numeric	Pool area in square feet
PoolQC	Categorical	Pool quality
Fence	Categorical	Fence quality
MiscFeature	Categorical	Miscellaneous feature not covered in other categories
MiscVal	Numeric	$Value of miscellaneous feature
MoSold	Numeric	Month Sold (MM)
YrSold	Numeric	Year Sold (YYYY)
SaleType	Categorical	Type of sale
SaleCondition	Categorical	Condition of sale

Techniques to be Used

We have decided to use **Artificial Neural Network** to predict the house prices. A lot of feature engineering and data cleaning is required to make the dataset suitable for Artificial Neural Network. The following are the packages to be used

Packages to be Used

Classifier	Package	Function
ANN	neuralnet, nnet, mlbench	neuralnet, nnet, train, traincontrol
Creating Folds	caret	createFolds
Area Under RoC Curve	AUC	auc
Converting categorical attributes into boolean attributes	ade4	acm.disjonctif
Computing and Plotting Correlations	corrplot	cor,corrplot

As and when the project proceeds, other packages will be included.

Experimental Methodology

As mentioned earlier, the project includes 3 phases namely,

1. Feature Engineering
2. ANN Training which involves parameter tuning, validation and evaluation
3. Comparison with established results and tuning the model for best results

Feature Engineering

Features play an important role in predictive models. Even a complex model might perform miserably when the dataset contains unnecessary features. We have decided to analyze the dataset and work on the following

1. Analysis of attributes
2. Converting categorical attributes to boolean attributes
3. Removing attributes which have a high correlation with each other
4. Removing attributes which have a low correlation with the output
5. Handling features with missing values
6. Handling features which have single value
7. Scaling and normalizing data

We are currently in this phase of the project.

Model Training

Training the ANN concerns with identifying the best set of parameters. Since it is a regression problem, the SalePrice to be predicted can be normalized to a value between 1 and 0 by dividing the values with the maximum value present and then multiplying the predicted value with the max value to get the price prediction. The following are the planned experimentations with parameters

Parameter Tuning and Resampling Techniques Planned

- K-fold cross validation - Experimenting with **different K values**
- Bootstrap
- Experimenting with the **number of hidden layers**
- Experimenting with the **number of nodes in every hidden layer**
- Experimenting with **functions available for error calculation**
- Experimenting with **available algorithms like backpropagation, rpropagation**.
- Experimenting with **different values for learning rate**
- Experimenting with **different threshold values**

Evaluation and Comparisons

The following evaluation techniques have been planned to be used

1. Accuracy

$$\frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

2. Precision

$$\frac{True\ Positive}{True\ Positive + False\ Positive}$$

3. Recall

$$\frac{True\ Positive}{True\ Positive + False\ Negative}$$

4. RMSE
5. Area under the RoC

Coding/Technology to be used

Programming Language	R
IDE	R Studio

Preliminary Results

We are currently in the **Feature Engineering** phase. The following sub-phases have been completed

Analysis of Attributes
The dataset contains categorical and numerical data.

Output - SalePrice	Numeric
No. of categorical attributes	46
No. of numerical attributes	34(Including output)

Handling Attributes with Single Values
Attributes with single values do not in anyway affect the output. Hence they are removed. R code to identify and eliminate single values has been completed. However the dataset does not contain any attribute with just single values

Handling Attributes with Missing Values
Attributes with missing values have been handled. Categorical attributes having missing values will be handled when the categorical attributes are converted into boolean attributes. However numerical attributes cannot have missing values. There are a number of ways to handle missing values in numeric attributes. Such attributes can either be removed, missing values can be substituted by zero. However the best possible way to handle numeric attributes is by substituting missing values with the mean of all the values held by the attribute.
The following are the numeric attributes with missing values.

1. LotFrontage
2. MasVnrArea
3. GarageYrBlt

GarageYrBlt is a year and we are researching if replacing the missing values in this column with the mean value makes sense. However the other two features have means as follows

1. LotFrontage - 70.04996

2. MasVnrArea-103.6853

Converting Categorical Attributes into Boolean Attributes
There are 46 categorical attributes. These attributes cannot be used directly to train ANN. Therefore, they must be converted to numerical attributes. To do this conversion, each possible value of the categorical variable is converted to a new boolean attribute. For example, the attribute 'poutcome' can take on one of the values in the set {'failure', 'nonexistent', 'success'}. This attribute is converted into three boolean attributes namely 'poutcome.failure', 'poutcome.nonexistent' and 'poutcome.success'. Each row will have exactly one of these three attributes set to '1' and the other two will be set to '0'.

Using this strategy,the number of attribute changed from **79 to 319**.

R code for the above has been attached along with the report.

Console Log

Reading the dataset from dataset/train.csv

Total Number of Attributes (Including Output) :: 80

Number of Categorical Columns :: 46

Categorical columns are :: MSSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual OverallCond RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition

Number of Numerical Columns (Including Output) :: 34

Categorical columns are :: LotFrontage LotArea YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SalePrice

Checking and Removing Attributes with Single Values

No Attribute contains Single Values

Correcting the following numeric attributes as they contain missing values: LotFrontage MasVnrArea GarageYrBlt

Updating Mean 70.04996 for missing values in feature LotFrontage

Updating Mean 103.6853 for missing values in feature MasVnrArea

Converting categorical attributes to boolean attributes

Number of attributes before conversion = 79

Number of attributes after conversion = 319