University of California, Los Angeles

Stats 101C: Lec 1

Introduction to Statistical Models and Data Mining

Predicting Affordability of Houses in Ames, Iowa

Author:
Yinsheng Wang, Xuelan Fu, Janice Li, Amanda Xu

December 14, 2018

Contents

1 Introduction

The main question we are going to address in this paper is: an appropriate way to predict the affordability of a housing given its various characteristics.

We approach the question first by data cleaning, which includes NA imputation and variable transformation. Second, we try different methods with R and they are Logistic Regression, KNN, SVM and Random Forest. Third we compare the results of each method and choose the one with highest accuracy rate.

2 Data Cleaning

2.1 Variables used to predict

(indicates that the variable is made by ourselves or has been changed from the original one)*

Table 1: Variables we use to predict

Variable Chart		
Variable Names	Type of Data	Description and Source
affordabilitty	Factor	Two level factor: Affordable and Unaffordable; Source: affordabilitty. This is the variable we want to predict.
MSZoning	Factor	Identifies the general zoning classification of the sale; Source: MSZoning.
LotArea	Numeric	Lot size in square feet; Source: LotArea.
Neighborhood	Factor	Physical locations within Ames city limits; Source: Neighborhood of given data.
BldgType	Factor	Type of dwelling; Source: BldgType.
OverallQual	Numeric	Scores of the overall material and finish of the house; Source: OverallQual.
OverallCond	Numeric	Scores of the overall condition of the house; Source: OverallCond

Variable Chart

Variable Names	Type of Data	Description and Source
SaleCondition	Factor	Condition of sale;Source: SaleCondition
RemodYN*	Numeric	Dummy variable (0-1) indicating whether the house has been remodeled or not; Source: YearBuilt, YearRemodAdd
Age*	Numeric	The age of the house; Source: YearBuilt.
GrLivArea	Numeric	Above grade (ground) living area square feet; Source: GrLivArea.
BsmtScore*	Numeric	Scores of basement according to its performance in height, general condition, walkout level, rating of basement finished area and rating of basement finished area (if multiple types). Source: BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2.
BsmtFSF*	Numeric	Finished basement square feet (sum of both type 1 and type 2); Source: BsmtFinSF1, BsmtFinSF2.
BsmtUnfSF	Numeric	Unfinished basement square feet. Source: BsmtUnfSF.
GarageScore*	Numeric	Scores of garage according to its performance in garage quality and condition and interior finish extent. Source: GarageFinish, GarageQual, GarageCond.
GarageCars	Numeric	Size of garage in car capacity; Source: GarageCars.
Gtype*	Numeric	Garage location; Source: GarageType
GarageAge*	Numeric	Age of garage; Source: GarageYrBlt.
FireplacesNum	Numeric	Number of fireplaces; Source: Fireplaces.
FireQu*	Numeric	Scores of fireplaces quality; Source: FireplaceQu.
X1stFarea	Numeric	First floor square feet; Source: 1stFlrSF.

Variable Chart

Variable Names	Type of Data	Description and Source
X2ndFarea	Numeric	Second floor square feet; Source: 2ndFlrSF.
MoSold	Numeric	Month sold; Source: MoSold.
YearSold	Numeric	Year sold; Source: YrSold.
BsmtFullBath	Numeric	Number of basement full bathrooms; Source: BsmtFullBath.
BsmtHalfBath	Numeric	Number of basement half bathrooms; Source: BsmtHalfBath.
FullBath	Numeric	Number of full bathrooms above grade; Source: FullBath.
HalfBath	Numeric	Number of half bathrooms above grade; Source: HalfBath.
Kitchen	Numeric	Number of kitchens above grade; Source: Kitchen.
KitchenQual*	Numeric	Scores of kitchen quality; Source: KitchenQual.
HouseStyle*	Numeric	Styles of dwelling; Source: HouseStyle.
MiscFeature*	Numeric	Dummy variable (0-1) indicating whether there is miscellaneous feature not covered in other categories; Source: MiscFeature.
TotalArea*	Numeric	Total dwelling area measured in square feet; Source: LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, GrLivArea, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, LowQualFinSF, PoolArea.
OpenPorchYN*	Numeric	Dummy variable (0-1) indicating whether there is open porch area; Source: OpenPorchSF.
EnclosedPorch*	Numeric	Dummy variable (0-1) indicating whether there is enclosed porch area; Source: EnclosedPorch.

Variable Chart

Variable Names	Type of Data	Description and Source
X3SsnPorch*	Numeric	Dummy variable (0-1) indicating whether there is three season porch area; Source: 3SsnPorch.
ScreanPorch*	Numeric	Dummy variable (0-1) indicating whether there is screen porch area; Source: ScreenPorch.
LotFrontage	Numeric	Linear feet of street connected to property; Source: LotFrontage.
Alley*	Numeric	Dummy variable (0-1) indicating whether there is alley access to property; Sourece: Alley.
LotShape*	Numeric	General shape of property; Source: LotShape
LandSlope*	Numeric	Slope of property; Source: LandSlope.
LandContour	Factor	Flatness of the property; Source: LandContour.
LotConfig	Factor	Lot configuration; Source: LotConfig.
Condition*	Numeric	Variable (0,1,2) indicating the proximity of property to various conditions; Source: Condition1, Condition2.
RoofStyle	Factor	Type of roof; Source: RoofStyle.
RoofMatl*	Numeric	Roof material; Source: RoofMatl.
Exterior1st	Factor	Exterior covering on house; Source: Exterior1st.
Exteriormore*	Numeric	Dummy variable (0-1) indicating whether there are more than one materials of exterior covering on house; Source: Exterior1st, Exterior2nd.
MasVnrType	Factor	Masonry veneer type; Source: MasVnrType.
Foundation	Factor	Type of foundation; Source: Foundation.

2.2 Transformation of variables

Table 2: Transformation of variables with * in part2.1

Transformation of variables with *		
Variable Names	Source Variable	Transformation
RemodYN*	YearBuilt YearRemodAdd	0 if YearBuilt and YearRemodAdd are the same; 1 if YearBuilt and YearRemodAdd are different.
Age*	YearBuilt	2010-YearBuilt.
BsmtScore*	BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2	First, change the factor of each variable into numeric, e.g. NA is 0 while Po is 1as the score of each property; Second, add scores for each variable together as the final score of basement.
BsmtFSF*	BsmtFinSF1 BsmtFinSF2	The sum of the value of two variables.
GarageScore*	GarageFinish GarageQual GarageCond	First, change the factor of each variable into numeric, e.g. NA is 0 while Po is 1; Second, add scores for each variable together as the final score of garage.
Gtype*	GarageType	Change the factor into numeric; NA: 0; Detchd and CarPort: 1; BuiltIn, Basment and Attchd: 2; 2Types: 3.
GarageAge*	GarageYrBlt	2010-GarageYrBlt.
FireQu*	FireplaceQu	Change the factor into numeric; NA: 0; Po: 1; Fa: 2; TA: 3; Gd: 4; Ex:5.

Transformation of variables with *

Variable Names	Source Variable	Transformation
KitchenQual*	KitchenQual	Change the factor into numeric; Po: 1; Fa: 2; TA: 3; Gd: 4; Ex: 5.
HouseStyle*	HouseStyle	1Story, SFoyer, SLyl: 1; 1.5Fin, 1.5Unf: 1.5; 2Story: 2; 2.5Fin, 2.5Unf: 2.5.
MiscFeature*	MiscFeature	NA: 0; TenC, Shed, Othr, Gar2, Elev: 1
TotalArea*	LofFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF 1stFlrSF 2ndFlrSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch	Sum of all the values of numeric variables listed.
OpenPorchYN*	OpenPorchSF	> 0: 1; $= 0$: 0.
EnclosedPorch*	EnclosedPorch	> 0: 1; $= 0$: 0.
X3SsnPorch*	3SsnPorch	> 0: 1; $= 0$: 0.
ScreanPorch*	ScreanPorch	> 0: 1; $= 0$: 0.

Transformation of variables with *

Variable Names	Source Variable	Transformation
Alley*	Alley	NA: 0; Grvl, Pave: 1.
LotShape*	LotShape	Reg: 0; IR1: 1; IR2: 2; IR3: 3.
LandSlope*	LandSlope	Gtl: 1; Mod: 2; Sev: 3.
Condition*	Condition1 Condition2	Condition1=Norm: 0; Condition1≠Norm, Condition2=Norm: 1; Condition2≠Norm: 2.
RoofMatl*	RoofMtl	Membran, WdShake, WdShng: 1; ClyTile, CompShg, Metal, Roll, Tar and Gry: 0.
Exteriormore*	Exterior1st Exterior2nd	Exterior1st=Exterior2nd: 0; Exterior1stExterior2nd: 1.

3 Methodology

- We try Logistic Regression first. But 49 variables are still too much for Logistic Regression, so we select a smaller and better subset using forward stepwise method, 10-fold CV serving as the selection standard.

- Second, we try KNN. But we have to decide the best K. After we decide the best K, which gives us the highest accuracy rate in 10-fold CV, we will use that K to predict.

- Third, comparing the results of Logistic Regression and KNN and the best K, we will see whether we need more flexible methods or less flexible one. If more flexible methods are suggested, we would try Random Forest and SVM. If less flexible methods are suggested, we would fit LDA and QDA model.

- Forth, we will choose the method giving highest accuracy rate in Kaggle leaderboard, and apply that model to get our prediction.

4 Main Results

The comparison of results of Logistic Regression and KNN indicates that a more flexible result is needed. So in total, we have tried Logistic Regression, KNN, SVM and Random Forest. And the accuracy rate of each method given by Kaggle is shown below:

Method	Accuracy Rate	Notice
Logistic Regression	0.63777	The best subset given by forward stepwise selection and 10-fold CV consists of MSZoing, Neighborhood, Age, GrLivArea, FireplacesNum. X1stFarea, FullBath, OpenPorchYN, RoofStyle and Foundation.
KNN	0.96000	The best K in our case is 1.
SVM	0.98000	Using tune() we choose the best parameters for SVM as cost=100, gamma=0.01.
Random Forest	0.98888	Set mtry from 1 to 49, and find the best one as 7.

Table 3: The comparison of results of four methods

The best method in our case is Random Forest, with accuracy rate of 0.988888 given by Kaggle leaderboard.

5 Limitations

- First, we have not compared the data from different rows. As suggested by professor, there are correlations between some rows.
- Second, there is still place for improvement in data cleaning and variable identification. Variables like Neighborhood play an important role

in deciding house affordability, while they are beyond our ability. Correct ways to deal with those variables may lead to higher accuracy.

- Third, the methods we use are those we have learnt in Stats 101C. A more flexible or a more advanced method may lead to better results.

6 Recommendation

- First, we suggest better and more careful way of data cleaning. We have 80 variables and every variable contains information about house affordability. We have to get a smaller subset while avoid losing information.

- Second, flexible methods are encouraged in model fitting and prediction.

7 Reference

James, Witten, Hastie, Tibshrani, "An Introduction to Statistical Learning with applications in R".