Minh Nguyen - Caine Li

Professor Steven Bogaerts

CSC 396

November 11th, 2019

Final Project Report

* Step 1: Study the Dataset, understand the goal.
* Step 2: Preprocessing the Data (Non-Numeric)
* I decided to handle non-numeric values first
* I add each, one by one to the predictors and see CV Average Score increase or decrease. Keep those that increase. (red = increase in CV Average Score)
* predictors = ['1stFlrSF','2ndFlrSF']: 0.5647996361440994
* predictors = ['1stFlrSF','2ndFlrSF','LotShape']: 0.5747607732337301
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','MSZoning']: 0.5738588489019477
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street']: 0.5748471405898814
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley']: 0.575893856546779
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour']: 0.5759009761866164
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Utilities']: 0.575889300973973
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'LotConfig']: 0.5749212840597682
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'LandSlope']: 0.5755863646671553
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood']: 0.5827429136663168
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1']: 0.5909597319161028
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2']: 0.5926051871675238
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','BldgType']: 0.5919439003751262
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle']: 0.5982225711518336
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'RoofStyle']: 0.5980813332679584
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'RoofMatl']: 0.5975082095551965
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st']: 0.6164743118001323
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd']: 0.6210931234245859
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'MasVnrType']: 0.6206523232438215
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual']: 0.623548743200765
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual', 'ExterCond']: 0.6249660741323005
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation']: 0.6608964869549114
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual']: 0.6650078542309423
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond' ]: 0.6728483245040247
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure' ]: 0.6804126584268262
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1' ]: 0.6998322774152279
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','BsmtFinType2' ]: 0.6993180683633995
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','Heating' ]: 0.6995699318844121
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC' ]: 0.7090479589092504
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir' ]: 0.7101157290572616
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical' ]: 0.7107488009249819
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual' ]: 0.7134994469431873
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional' ]: 0.7200267444954143
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu' ]: 0.7248109113381143
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType' ]: 0.7286475702434075
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish' ]: 0.7291751717922148
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageQual' ]: 0.7290582884566005
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond' ]: 0.7297848076054614
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond','PavedDrive' ]: 0.7307585635149326
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond','PavedDrive','PoolQC' ]: 0.7300515712083857
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond','PavedDrive','Fence' ]: 0.7306366940125868
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond','PavedDrive','MiscFeature' ]: 0.7302853917379489
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond','PavedDrive','SaleType' ]: 0.731558979677952
* predictors = ['1stFlrSF','2ndFlrSF','LotShape','Street','Alley', 'LandContour', 'Neighborhood', 'Condition1', 'Condition2','HouseStyle', 'Exterior1st', 'Exterior2nd', 'ExterQual','ExterCond', 'Foundation','BsmtQual','BsmtCond','BsmtExposure', 'BsmtFinType1','HeatingQC','CentralAir','Electrical','KitchenQual', 'Functional', 'FireplaceQu','GarageType', 'GarageFinish', 'GarageCond','PavedDrive','SaleType', 'SaleCondition' ]: 0.7309687249440768
* Step 3: Preprocessing data (Numeric)
  + GarageYrBlt: median because no decimals and no chance of outliers
  + LotFrontage: mean for average length
  + MasVnrArea: mean for average length
  + GarageArea & GarageCars: fillna(0) since Na means no garage.
  + Standardize numeric data to put all numeric data on an equal scale
* Step 4: Reconsider predictors
  + After reading your comment on our checkpoint about how we preprocess non-numeric data, we use pd.get\_dummies to handle all attributes that we believe to be nominal.
  + We then again added each attribute one by one (only those that used get\_dummies and the new numerics) to get higher CV Score.
  + Redo preprocess data with nominal attributes (attributes that we converted to ordinal before) and test them add them in predictors one more time to see if they increase CV Score. Keep those that increase the CV Score.
  + CV Score up until this point: CV Average Score: 0.8069438247115845
  + Add numerics values into predictors one by one. Keep ones that increase CV Score.
  + CV Score up until this point: CV Average Score: 0.8398417095283461
  + Created dummies for ExterQual
  + CV Score up until this point: CV Average Score: 0.8483510070822161
* Step 5: Creating Models
  + Used builtin KNeighborsClassifier and got error: The least populated class in y has only 1 members, which is too few. The minimum number of members in any class cannot be less than n\_splits=20.
  + CV Average Score: 0.02321052515844024
  + Used decision tree
  + Accuracy Score on train data: 0.9990867579908675
  + Accuracy Score on test data: 0.010958904109589041
  + We can’t figure out why accuracy for KNN and decision tree is so low. Syntax seems to match previous projects and messing with transformData didn’t achieve anything.
  + By changing scoring =’r2’, n\_neighbors = 1, and cv=2 the score increased ( Finally we figured it out)
  + CV Average Score for KNN: 0.6803068062178244
* Step 6: Creating output file to upload on kaggle
  + Create an output file for KNN (Name testResultsKNN)
  + Create an output file for decision tree model (Name testResutlsCLF)
* Submit the results.