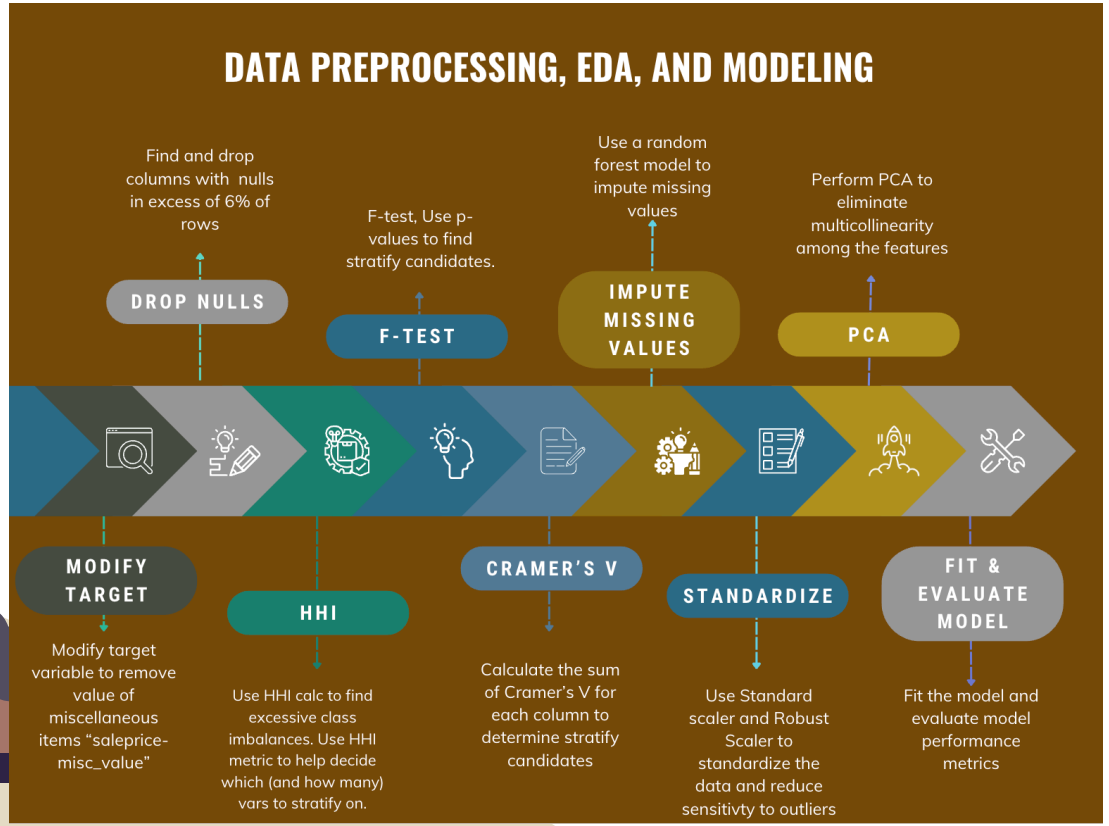
Housing sale price predictor

Nicholas Sanso

Objective

Create a model that will predict the sales price of housing in low and medium density residential areas

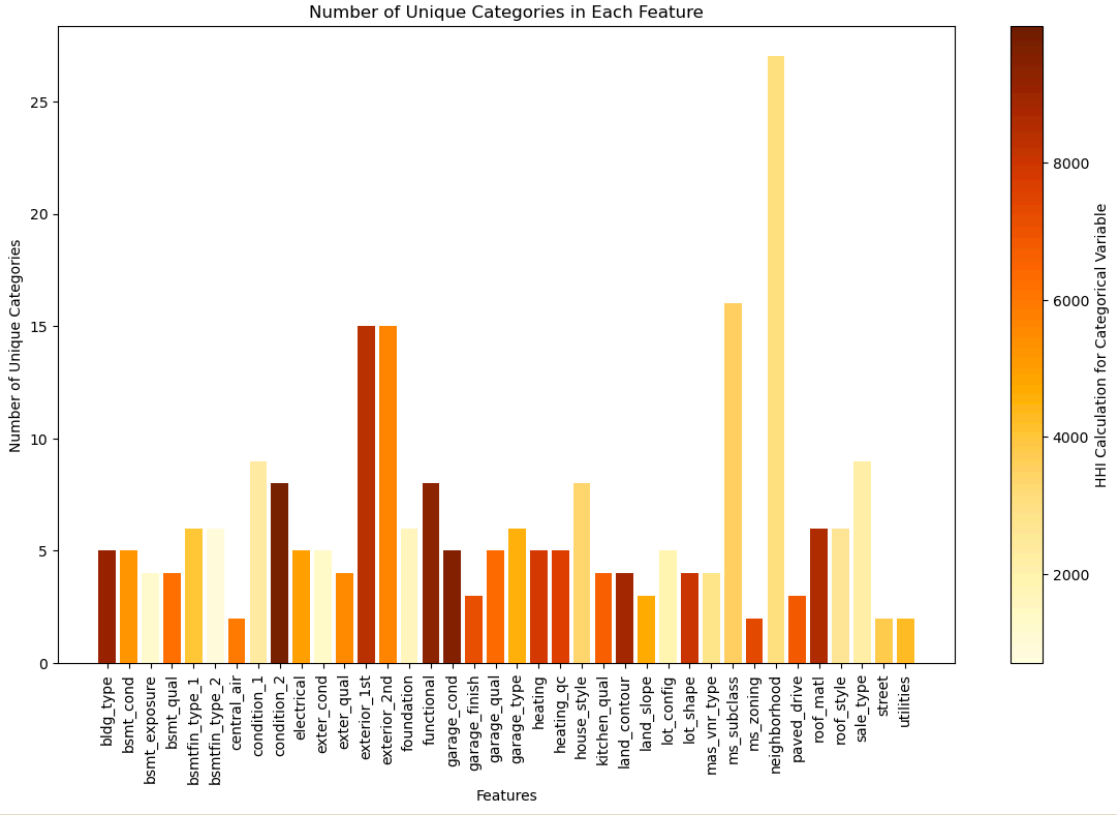
Workflow


DATA PREPROCESSING, EDA, AND MODELING
Find and drop columns with nulls in excess of 6% of rows
DROP NULLS
F-test, Use p-values to find stratify candidates.
F-TEST
Use a random forest model to impute missing values
IMPUTE MISSING VALUES
Perform PCA to eliminate multicollinearity among the features
PCA
MODIFY TARGET
Modify target variable to remove value of miscellaneous items "saleprice-misc_value"
HHI
Use HHI calc to find excessive class imbalances. Use HHI metric to help decide which (and how many) vars to stratify on.
CRAMER'S V
Calculate the sum of Cramer's V for each column to determine stratify candidates
STANDARDIZE
Use Standard scaler and Robust Scaler to standardize the data and reduce sensitivty to outliers
FIT & EVALUATE MODEL
Fit the model and evaluate model performance metrics

Categorical Concentration

Perform HHI calculations on the categories in the features.

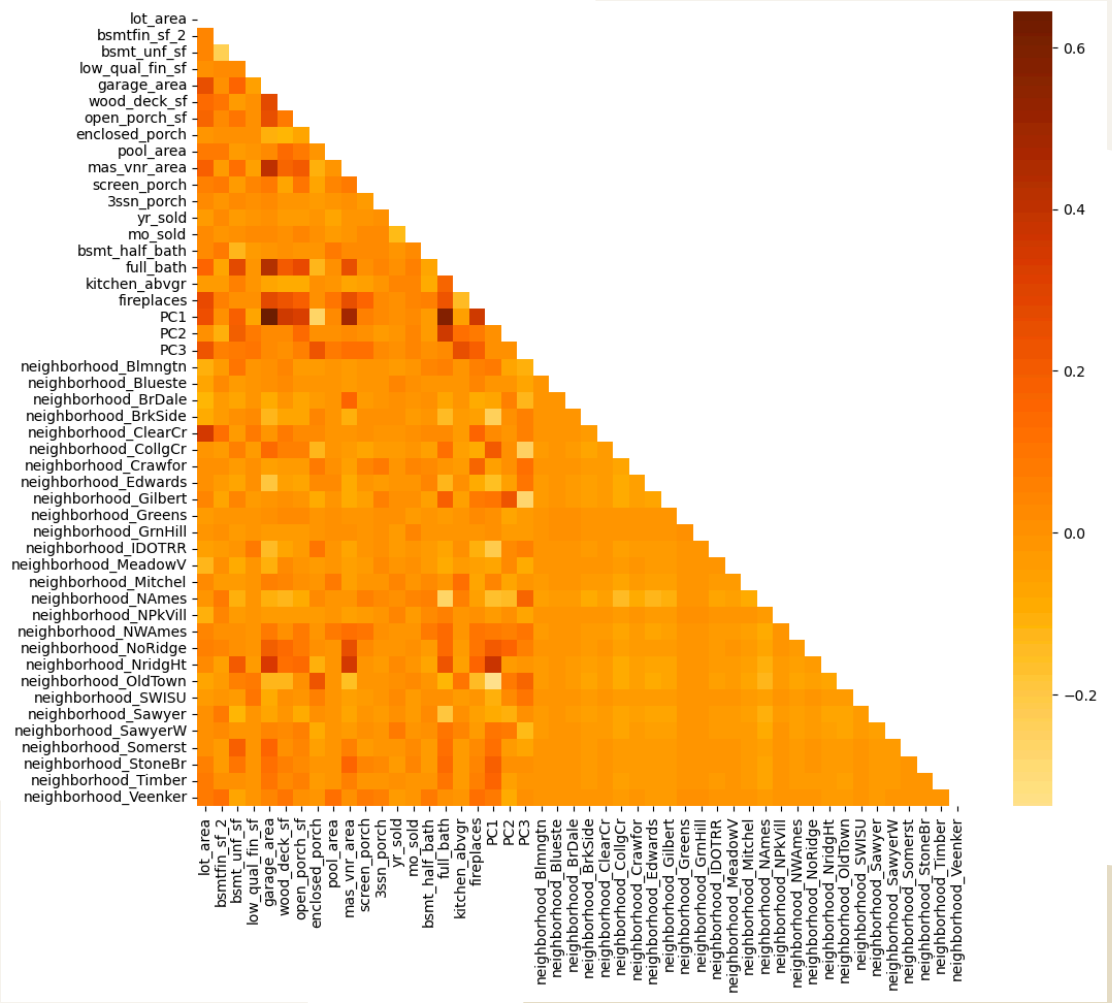
Exclude highly centralized categories which hinder ability to split observations into training and testing sets.



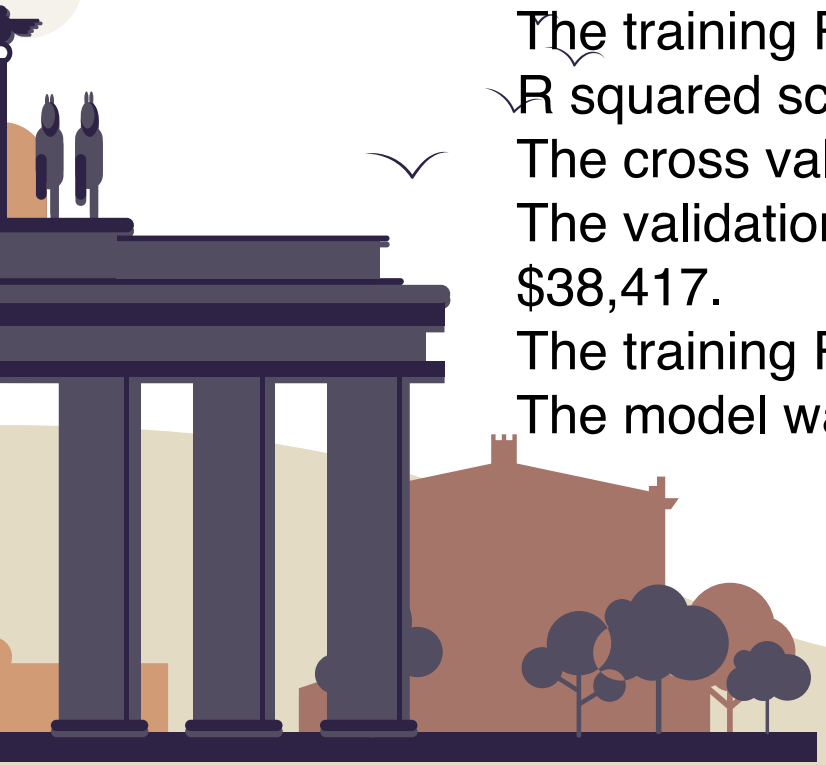
HEatMap

Heatmap after PCA showing the correlation between different features.

Correlations for variables included in the final model were low.



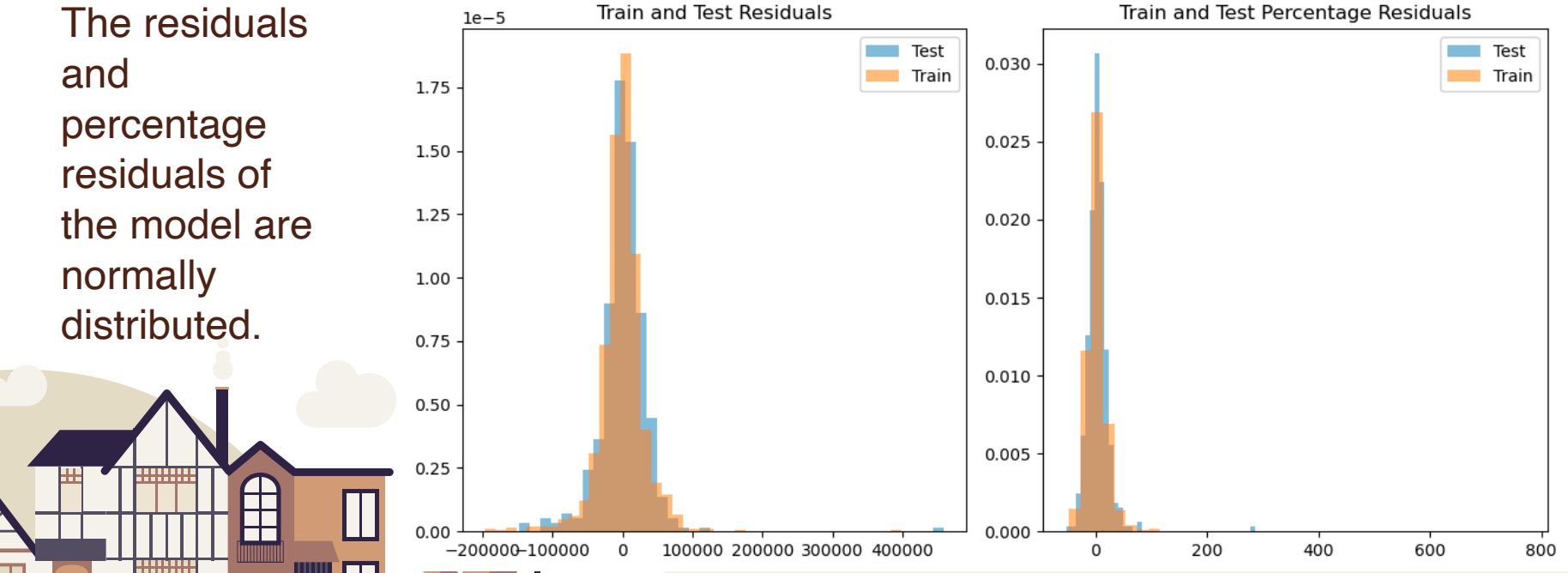
Model Results



The training R squared score was .838 and the test R squared score was .75.
The cross validation scores ranged from .73 to .83.
The validation Root Mean Square Error was $38,417.
The training Root Mean Square Error was $32,307.
The model was predictive and not over fit.

Evaluating the residuals

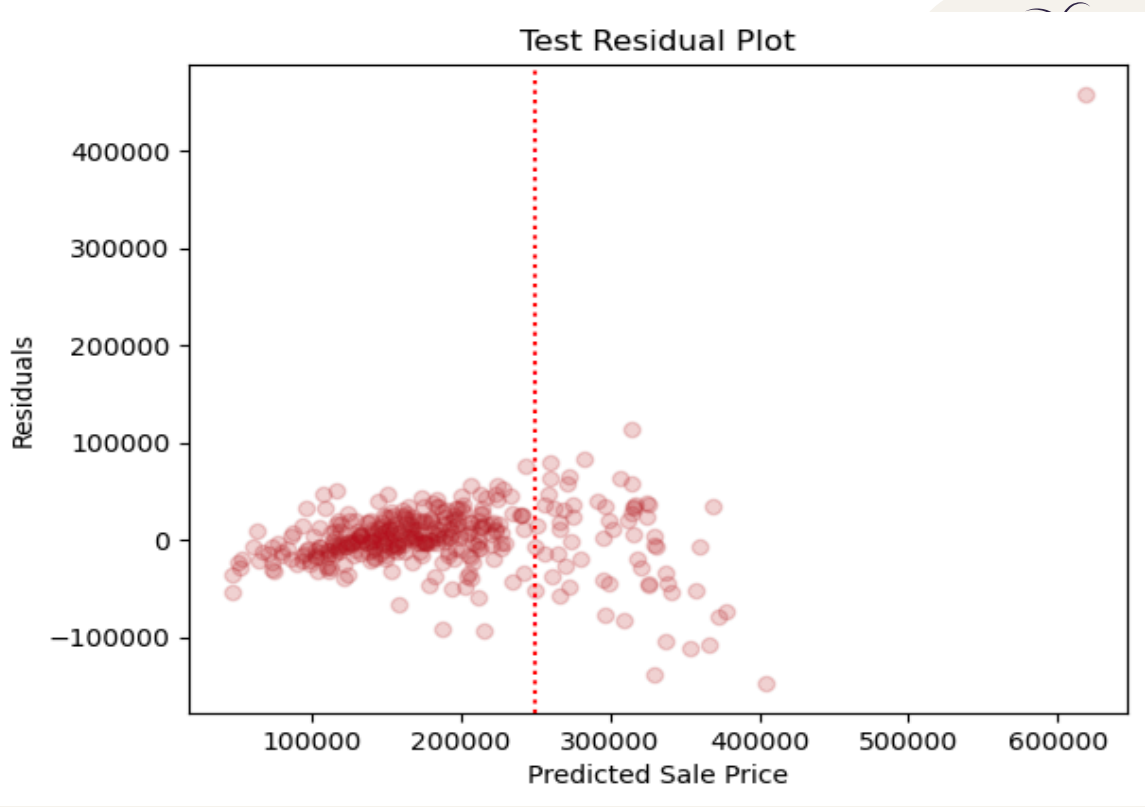
The residuals and percentage residuals of the model are normally distributed.



Evaluating the residuals

Heteroskedasticity appears in the residuals as the predicted salesprice increases.

The model should only be used for houses whose predicted sales price is less than $250,000.



Primary Conclusions

1. There are probably a limited number of patterns that the model relies heavily on.
2. The improvement in performance with the reduction in batch size indicates that the model may have been converging to a local minimum when minimizing the loss function during the (training) gradient descent process.
3. Considerations for model performance must be balanced against hardware requirements. MaxPooling2D's contribution to reducing training time is indispensable for all CNN models.
4. A model that outperformed the base model was established.