HOUSE SALE PRICE PREDICTION

SUBMITTED BY:

DEEPRO SENGUPTA

BACKGROUND

- Client is an United States based real estate firm
- The firm looking to expand into the Australian real estate market.

THE PROBLEM STATEMENT

- Identify which independent variables affect housing sale price the most.
- By what degree do each independent variables affect housing price.
- Build a model to predict housing sale price.

THE DATA

- Data source: Provided by Client
- Contains 81 columns and 1,168 rows
- Target Variable name: ‘SalePrice’

THE DATA: SUMMARY STATISTICS KEY TAKEAWAYS

- The mean is larger then the 50th percentile and there is a huge difference between the 75th percentile and max values for many columns. This indicates the presence of outliers.

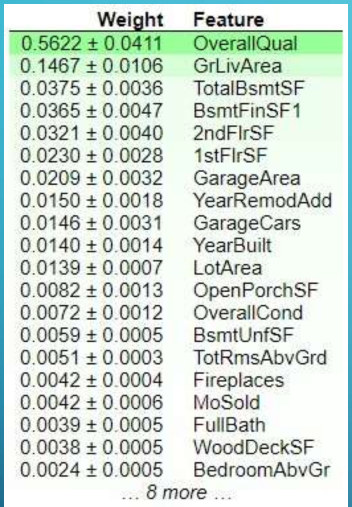
DATA PRE-PROCESSING STEPS

- Step 1: Handling Missing values:
  - Categorical Columns: All missing values replaced with mode of the column.
  - Numeric Columns: All missing values replaced using random number imputation.
- Step 2: Hypothesis Testing for Identifying Significant Independent Columns
  - Categorical columns: Chi-square test of significance.
  - Numeric columns: Pearson correlation test.
- Step 3: Skewness Removal: The following methods were used to handle skewness.
  - Inverse Transformation: on column MSSubClass.
  - Log Transformation: on column LotArea.
  - Cube-root Transformatino: on columns OverallCond, GrLivArea, BsmtFullBath, KitchenAbvGr, TotRmsAbvGrd, & Fireplaces
  - Square-root Transformation: on columns BsmtFinSF1, BsmtUnfSF, 1stFlrSF, & 2ndFlrSF
  - Yea-Johnson Transformation: on columns TotalBsmtSF, HalfBath, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea & YearBuilt.
- Step 4: Outlier Removal using IQR-Method.

DATA MODELLING & EVALUATION

- Algorithms used:
  - Linear Regression
  - Random Forest Regression
  - Decision Tree Regression
  - Support Vector Regression
  - K-Neighbors Regression
  - Multi-Layer Perceptron Regression
- Evaluation Metrics: R2-Score & RSME
- Final Model Selection Criterion:
  - It has the highest R2-score.
  - The has the lowest RSME indicating that the errors are not widely spread.
  - Test-Train score difference is not high indicating there is no overfitting.
- Final Model: Random Forest Regressor

CONCLUSION



Weight	Feature
0.5622 ± 0.0411	OverallQual
0.1467 ± 0.0106	GrLivArea
0.0375 ± 0.0036	TotalBsmtSF
0.0365 ± 0.0047	BsmtFinSF1
0.0321 ± 0.0040	2ndFlrSF
0.0230 ± 0.0028	1stFlrSF
0.0209 ± 0.0032	GarageArea
0.0150 ± 0.0018	YearRemodAdd
0.0146 ± 0.0031	GarageCars
0.0140 ± 0.0014	YearBuilt
0.0139 ± 0.0007	LotArea
0.0082 ± 0.0013	OpenPorchSF
0.0072 ± 0.0012	OverallCond
0.0059 ± 0.0005	BsmtUnfSF
0.0051 ± 0.0003	TotRmsAbvGrd
0.0042 ± 0.0004	Fireplaces
0.0042 ± 0.0006	MoSold
0.0039 ± 0.0005	FullBath
0.0038 ± 0.0005	WoodDeckSF
0.0024 ± 0.0005	BedroomAbvGr
… 8 more …	

- The table on left shows results after performing Permutation importance.
  - While only significant columns were considered for model building, permutation importance reveals OverallQual is the most important feature.
- R2-score of the final model is 83.19%
- RSME of the final model is 33500.74