

Project 3: House Price Prediction

Group Alan Turing

8 November 2024



Meet the Team



DENDI SUNARDI

Your Role/Contribution

[dendisunardi](#)



MISBAHUL MUNIR

End-to-End Modelling and Analysis

[Misbahul Munir](#)



Chelsea Castro

Dataset Preparation & Initial Analysis

[Chelsea Castro](#)



Rio Djajoesman

Future Improvement

[riodjajoesman](#)



Teddy Budiwan

Your Role/Contribution

[teddybudiwan](#)



Ismail

Your Role/Contribution

[Linkedin Name](#)



HINU HARDHANTO

Model quality control

[hinu hardhanto](#)

Content

- Background
- Problem Statement
- Objective & Scope
- Data Collection & Preparation
- Model Development
- Training & Optimization
- Results
- Real-world Application
- Future Improvement
- Conclusion



Background



Pasar real estat memainkan peran penting dalam ekonomi global, memengaruhi stabilitas keuangan, tren investasi, dan kekayaan pribadi. Memprediksi harga rumah secara akurat merupakan tantangan rumit yang memengaruhi berbagai pemangku kepentingan, termasuk pembeli, penjual, agen real estat, dan pembuat kebijakan. Harga rumah dipengaruhi oleh berbagai faktor, seperti fitur properti, lokasi, kondisi ekonomi, permintaan pasar, dan lain-lain

Secara tradisional, penilaian rumah bergantung pada penilai ahli yang mempertimbangkan faktor kualitatif dan kuantitatif. Namun, kemajuan dalam ilmu data dan machine learning telah membuka jalan bagi pendekatan yang lebih sistematis dan berbasis data. Dengan memanfaatkan data historis dan algoritma machine learning, kita dapat membangun model prediktif yang mengidentifikasi pola dan tren dalam data perumahan, yang mengarah pada prediksi harga yang lebih andal dan objektif.

Background

Latar belakang lainnya yang mendorong perlunya sebuah sistem house price prediction adalah diantaranya:

- **Fluktuasi Pasar yang Dinamis:**
 - Machine learning membantu adaptasi terhadap perubahan cepat di pasar properti.
- **Akurasi Penilaian Properti:**
 - Prediksi harga yang akurat mempercepat transaksi jual beli.
- **Pengurangan Bias:**
 - Analisis berbasis data mengurangi subjektivitas dalam penentuan harga.
- **Aksesibilitas Bagi Pembeli & Investor:**
 - Sistem real-time memudahkan keputusan pembelian berbasis data.
- **Optimasi Keputusan Investasi:**
 - Prediksi presisi mendukung identifikasi investasi menguntungkan.



Problem Statement

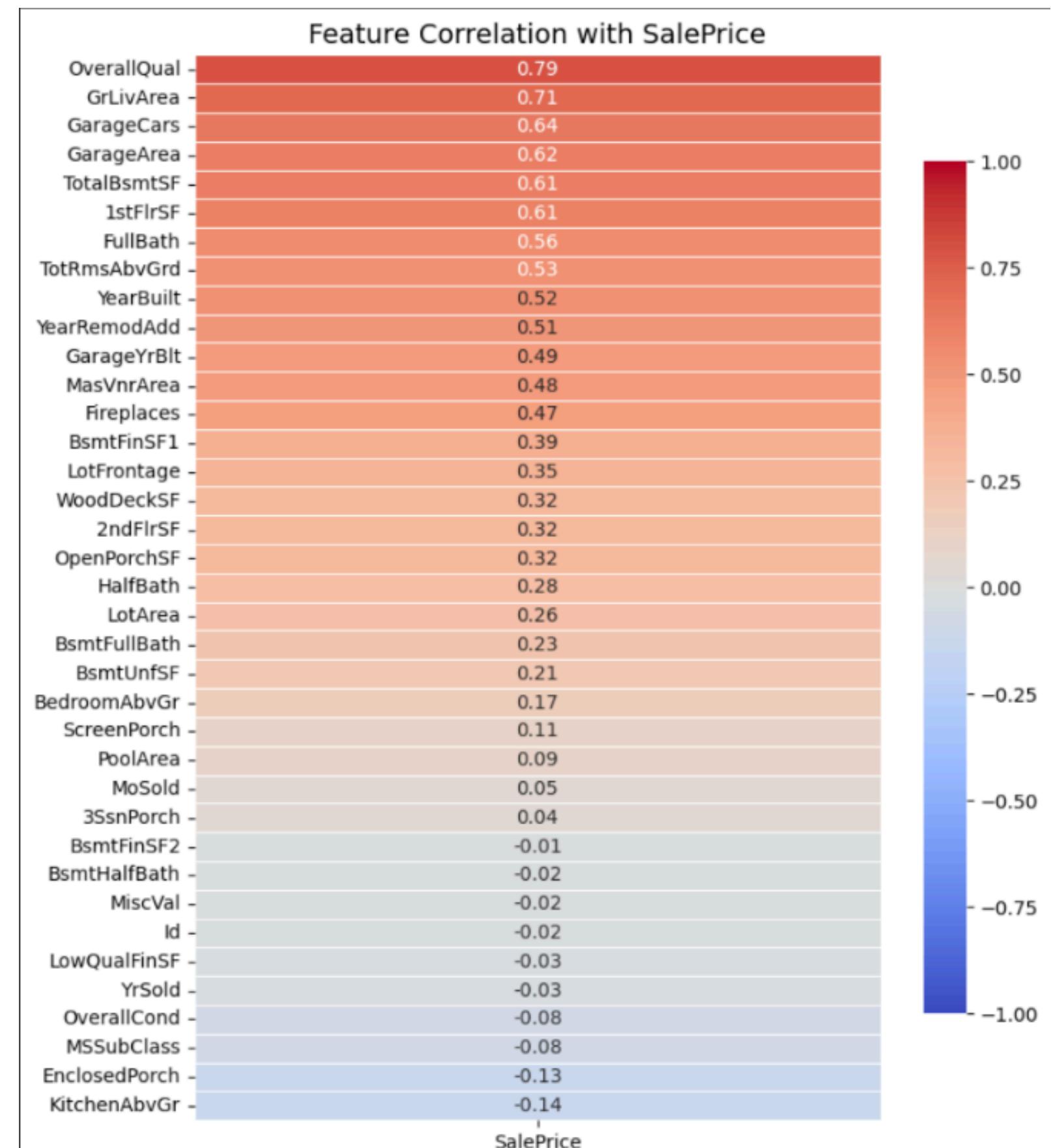
- Tujuan dari proyek ini adalah untuk mengembangkan model prediktif menggunakan *machine learning*, untuk memperkirakan harga rumah berdasarkan fitur-fitur utama seperti ukuran properti, lokasi, kualitas konstruksi, fasilitas, akses jalan dan lain-lain.
- Ketidakakuratan Estimasi Harga: Metode penilaian harga rumah tradisional sering tidak akurat dan kurang responsif terhadap perubahan pasar, menghambat pengambilan keputusan yang tepat bagi pembeli, penjual, dan investor.
- Keterbatasan Transparansi dan Efisiensi: Ketiadaan sistem prediksi berbasis data yang cepat dan transparan mengakibatkan proses penilaian yang lambat dan rentan terhadap bias, mengurangi kepercayaan dan efisiensi di pasar properti.
- Secara teknis diperlukan mengatasi tantangan pada mengidentifikasi prediktor yang paling penting dan menangani masalah seperti data yang hilang dan pemilihan fitur untuk membangun model yang dapat digeneralisasi dengan baik ke data yang tidak terlihat.

Objectives & Scope

- Tim Alan Turing mengimplementasikan beberapa model algoritma untuk memprediksi harga rumah dan melakukan optimasi algortima (hyperparameter tuning).
- Dari hasil eksperimen, lakukan evaluasi dan penarikan kesimpulan mana algoritma terbaik.

Data Collection & Preparation

- Dataset ini mungkin memiliki 1460 baris x 80 kolom, jumlah yang secara signifikan lebih sedikit dibandingkan dataset sebelumnya, tetapi tetap menantang karena setiap variabel memerlukan analisis khusus sebelum dapat ditentukan apakah akan dihapus atau tidak.
- Setiap variabel memiliki dampak yang berbeda terhadap SalePrice. Misalnya, variabel kolam renang yang memiliki hampir ~97% nilai kosong ternyata dapat meningkatkan harga jual secara signifikan, yang pada akhirnya dapat menyebabkan skew dalam data.
- Kami melakukan korelasi fitur awal antara variabel numerik dan SalePrice untuk mendapatkan pemahaman awal terhadap data, sebelum melakukan analisis mendalam terhadap setiap variabel untuk menentukan mana yang perlu dihapus dan mana yang bisa digabungkan.

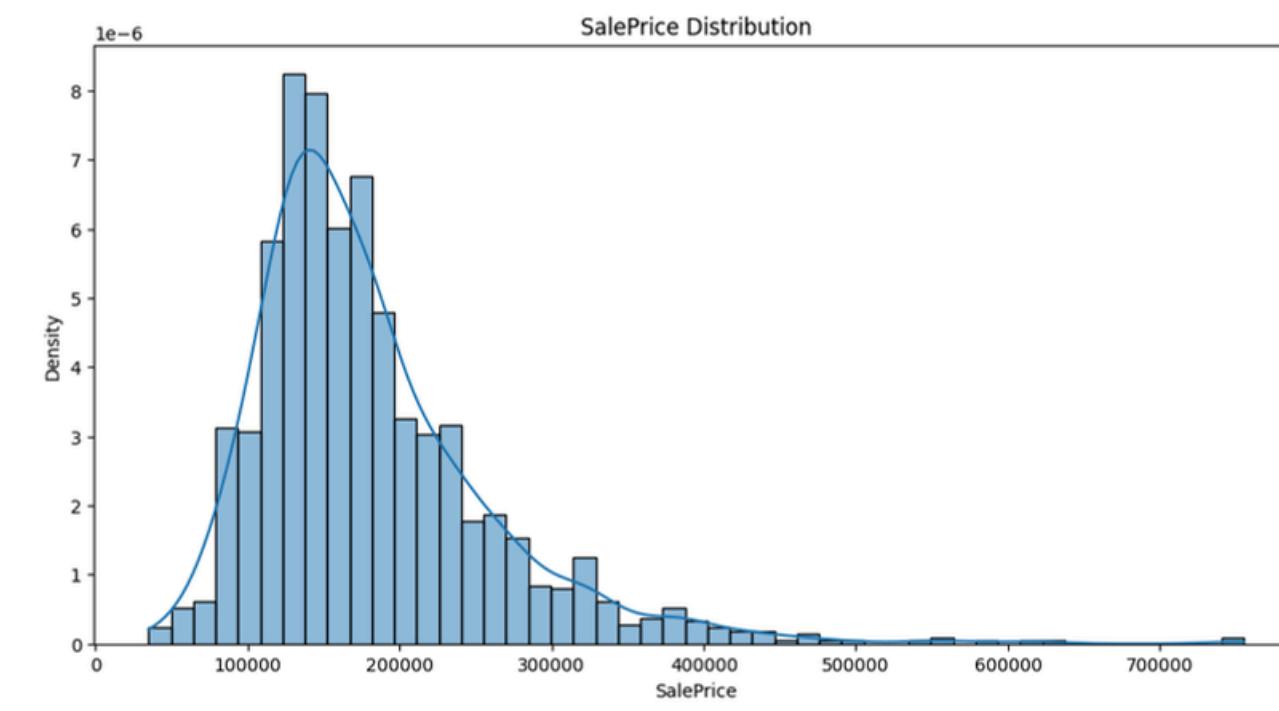
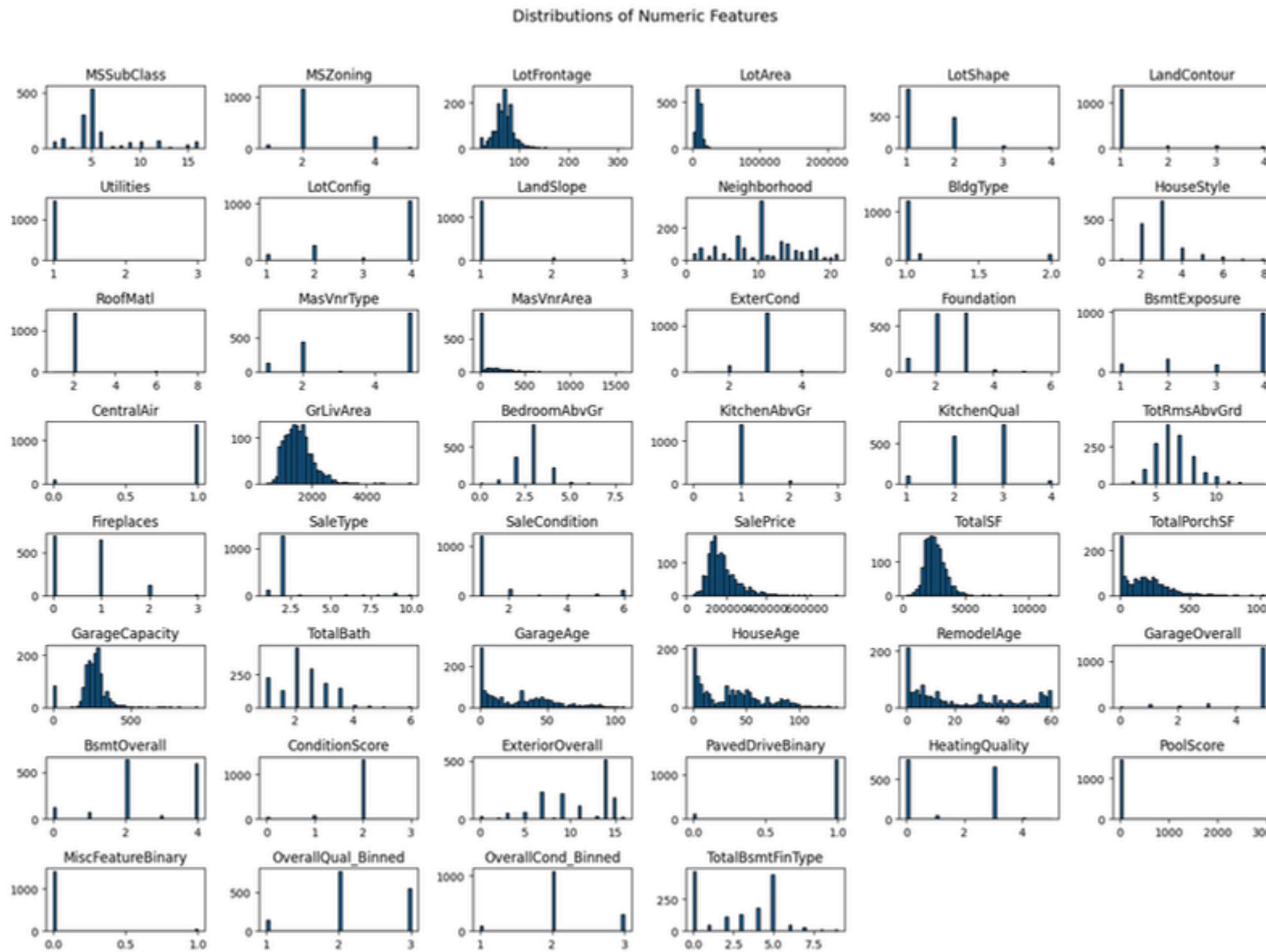


Data Collection & Preparation

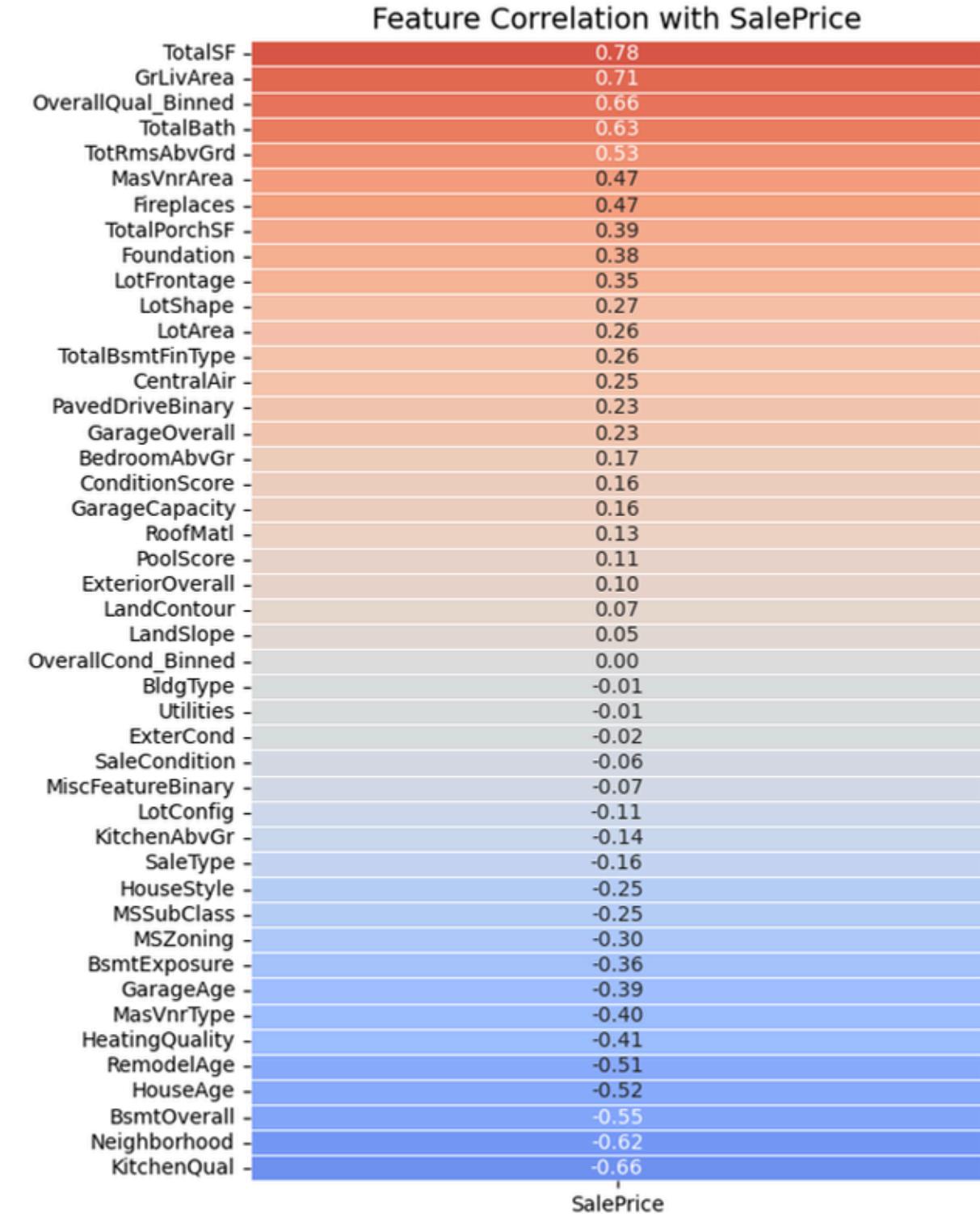
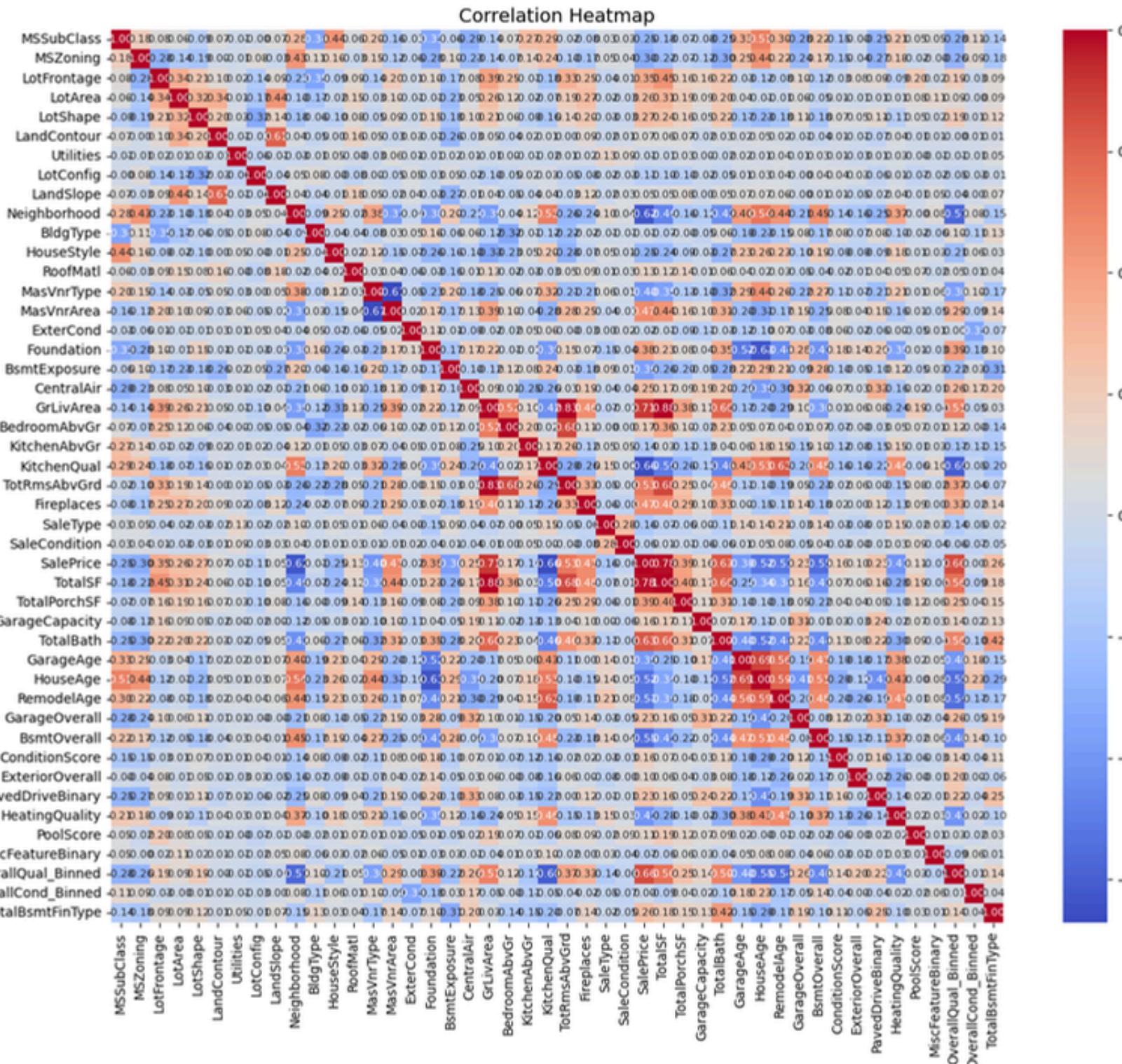
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 46 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   MSSubClass        1460 non-null   int64  
 1   MSZoning          1450 non-null   float64 
 2   LotFrontage       1460 non-null   float64 
 3   LotArea           1460 non-null   int64  
 4   LotShape          1460 non-null   int64  
 5   LandContour      1460 non-null   int64  
 6   Utilities         1460 non-null   int64  
 7   LotConfig         1456 non-null   float64 
 8   LandSlope         1460 non-null   int64  
 9   Neighborhood      1094 non-null   float64 
 10  BldgType          1334 non-null   float64 
 11  HouseStyle        1460 non-null   int64  
 12  RoofMatl          1460 non-null   int64  
 13  MasVnrType        1460 non-null   int64  
 14  MasVnrArea        1460 non-null   float64 
 15  ExterCond         1460 non-null   int64  
 16  Foundation        1460 non-null   int64  
 17  BsmtExposure     1460 non-null   int64  
 18  CentralAir        1460 non-null   int64  
 19  GrLivArea         1460 non-null   int64  
 20  BedroomAbvGr     1460 non-null   int64  
 21  KitchenAbvGr     1460 non-null   int64  
 22  KitchenQual       1460 non-null   int64  
 23  TotRmsAbvGrd     1460 non-null   int64  
 24  Fireplaces        1460 non-null   int64  
 25  SaleType          1460 non-null   int64  
 26  SaleCondition     1460 non-null   int64  
 27  SalePrice         1460 non-null   int64  
 28  TotalSF           1460 non-null   int64  
 29  TotalPorchSF      1460 non-null   int64  
 30  GarageCapacity    1460 non-null   float64 
 31  TotalBath          1460 non-null   float64 
 32  GarageAge          1460 non-null   float64 
 33  HouseAge           1460 non-null   int64  
 34  RemodelAge         1460 non-null   int64  
 35  GarageOverall      1460 non-null   int64  
 36  BsmtOverall        1460 non-null   int64  
 37  ConditionScore     1460 non-null   int64  
 38  ExteriorOverall    1460 non-null   int64  
 39  PavedDriveBinary   1460 non-null   int64  
 40  HeatingQuality     1460 non-null   int64  
 41  PoolScore          1460 non-null   float64 
 42  MiscFeatureBinary  1460 non-null   int64  
 43  OverallQual_Binned 1460 non-null   int64  
 44  OverallCond_Binned 1460 non-null   int64  
 45  TotalBsmtFinType   1460 non-null   float64 
dtypes: float64(11), int64(35)
```

- **Agregasi Fitur:** Membuat fitur baru untuk luas total, kapasitas garasi, jumlah kamar mandi, usia rumah, dan area teras dengan menggabungkan kolom-kolom terkait.
 - Menggabungkan atribut garasi dan basement dalam kolom baru seperti GarageCapacity, GarageAge, GarageOverall, dan BsmtOverall, lalu mengisi nilai kosong dengan angka 0 atau nilai mode.
 - Menggabungkan kolom Condition1 dan Condition2 menjadi ConditionScore, serta Exterior1st dan Exterior2nd menjadi ExteriorOverall, sambil mengisi nilai kosong.
 - Fitur Biner: Mengonversi PavedDrive dan MiscVal menjadi fitur biner, menunjukkan ada/tidaknya aspal di jalan masuk dan fitur lain.
 - **Mengisi Nilai Kosong:** Mengisi nilai kosong pada kolom LotFrontage berdasarkan median di setiap Neighborhood, serta mengisi kolom MasVnrType dan MasVnrArea.
 - **Fitur Biner:** Mengonversi PavedDrive dan MiscVal menjadi fitur biner, menunjukkan ada/tidaknya aspal di jalan masuk dan fitur lain.
 - **Skor Kualitas:** Membuat kolom PoolScore berdasarkan PoolArea dan PoolQC, dan mengkategorikan OverallQual dan OverallCond ke dalam kategori Rendah, Sedang, dan Tinggi.
 - **Peringkat Ordinal:** Menerapkan peringkat ordinal pada kolom kategori tertentu (seperti MSSubClass, MSZoning, LotConfig) menggunakan nilai integer.
 - **Label Encoding:** Menggunakan label encoding pada kolom ordinal (GarageOverall, BsmtOverall, ConditionScore, ExteriorOverall, HeatingQuality) untuk persiapan analisis.
 - **Penghapusan Kolom:** Menghapus kolom-kolom yang telah digabungkan ke dalam fitur baru atau yang tidak diperlukan dalam analisis lanjutan.

Data Collection & Preparation



Data Collection & Preparation



Model Development

Pengembangan Model dilakukan menggunakan beberapa Algoritma dan teknik Hyperparameter Tuning

1. Algoritma Model

- Linear Regression
- Random Forest
- XG Boost
- Ada Boost

2. Hyperparameter Tuning

- with Bayesian Optimization: pendekatan ini lebih efisien dan efektif

Training & Optimization

Bayesian Optimization digunakan untuk hyperparameter tuning karena pendekatan ini lebih efisien dan efektif dibandingkan dengan metode pencarian hyperparameter tradisional seperti grid search dan random search.

Bayesian Optimization

- Efisiensi dalam pencarian ruang Hyperparameter
- Lebih cepat dan lebih sedikit eksperimen yang diperlukan

Cross Validation

- Model diuji pada beberapa folds, sehingga hasil akhir lebih representatif dan generalisasi model menjadi lebih baik

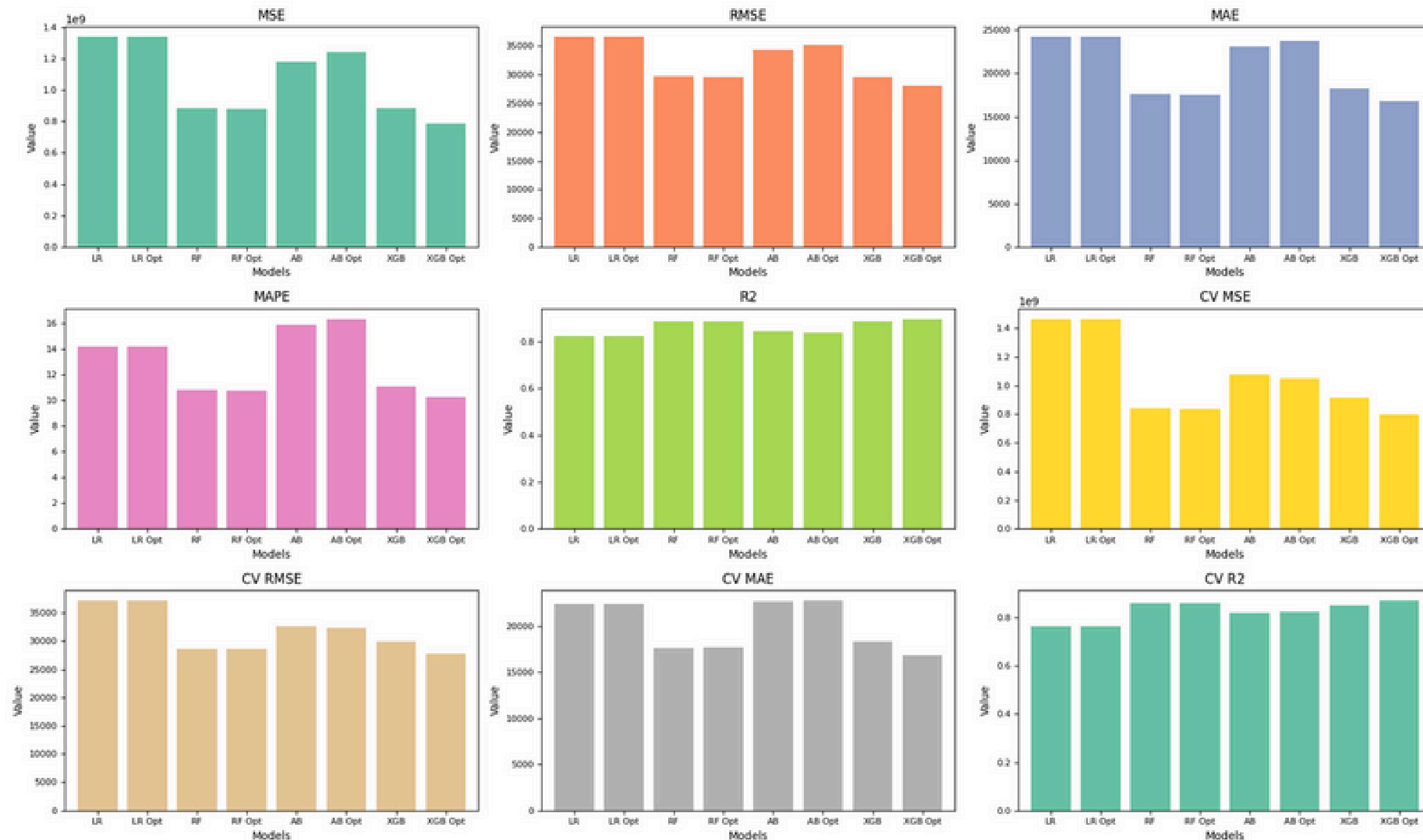
Training & Optimization

	LinearRegression Initial	LinearRegression Optimized	RandomForest Initial	RandomForest Optimized	AdaBoost Initial	AdaBoost Optimized	XGBoost Initial	XGBoost Optimized
MSE	1.335224e+09	1.335224e+09	8.8e+08	8.790621e+08	1.180081e+09	1.238830e+09	8.807470e+08	7.867161e+08
RMSE	3.654072e+04	3.654072e+04	2.9e+04	2.9e+04	3.435231e+04	3.519701e+04	2.967738e+04	2.804846e+04
MAE	2.415786e+04	2.415786e+04	1.764738e+04	1.756820e+04	2.304866e+04	2.370140e+04	1.827180e+04	1.684765e+04
MAPE	1.419415e+01	1.419415e+01	1.078014e+01	1.073398e+01	1.585412e+01	1.630375e+01	1.108232e+01	1.024358e+01
R2	8.259234e-01	8.259234e-01	8.848105e-01	8.853947e-01	8.461498e-01	8.384908e-01	8.851748e-01	8.974338e-01
CV MSE	1.462779e+09	1.462779e+09	8.398114e+08	8.374937e+08	1.077381e+09	1.049900e+09	9.116213e+08	7.974661e+08
CV RMSE	3.711834e+04	3.711834e+04	2.856768e+04	2.855333e+04	3.265014e+04	3.230351e+04	2.982634e+04	2.780105e+04
CV MAE	2.240536e+04	2.240536e+04	1.764105e+04	1.773371e+04	2.270673e+04	2.277896e+04	1.832130e+04	1.687226e+04
CV R2	7.642980e-01	7.642980e-01	8.601966e-01	8.607298e-01	8.185474e-01	8.214757e-01	8.492743e-01	8.690738e-01

- Model Linear Regression, baik dalam versi awal maupun yang sudah dioptimasi, memiliki performa yang serupa dengan MSE sekitar 1,33e+09 dan RMSE sekitar 3,65e+04. Nilai R2 (0,825) menunjukkan model ini cukup mampu menjelaskan variabilitas data, namun tidak mencapai performa terbaik dibandingkan model lainnya.
- Model Random Forest memiliki MSE yang jauh lebih rendah dibandingkan Linear Regression, yaitu sekitar 8,8e+08 untuk model awal dan 8,7e+08 setelah optimasi. Nilai R2-nya mencapai sekitar 0,85 yang menunjukkan peningkatan ketepatan prediksi. Model ini menunjukkan bahwa optimasi menghasilkan perbaikan performa yang kecil namun signifikan.
- AdaBoost, baik versi awal maupun versi yang sudah dioptimasi, memiliki performa MSE dan RMSE yang lebih tinggi dibandingkan Random Forest, dengan nilai MSE sekitar 1,18e+09 (awal) dan 1,23e+09 (setelah optimasi). R2-nya mencapai 0,84, sedikit lebih rendah dari Random Forest dan XGBoost.
- Model XGBoost menunjukkan performa terbaik dalam hal MSE dan RMSE, dengan nilai MSE sebesar 8,8e+08 untuk model awal dan meningkat menjadi 7,87e+08 setelah optimasi. R2 model ini yang dioptimasi mencapai sekitar 0,87, yang tertinggi di antara semua model.
- Model XGBoost yang telah dioptimasi adalah model terbaik berdasarkan nilai MSE, RMSE, dan R2, menunjukkan bahwa model ini paling efektif dalam memprediksi target dengan kesalahan yang paling rendah dan ketepatan prediksi yang paling tinggi.**

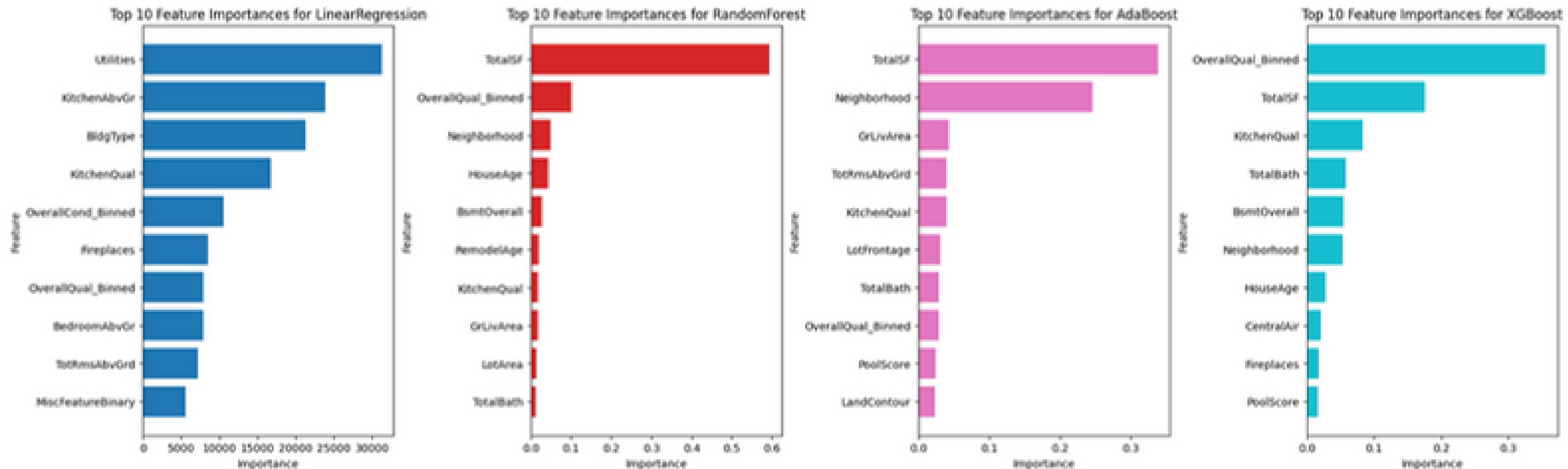
Results

Model Performance Metrics Comparison

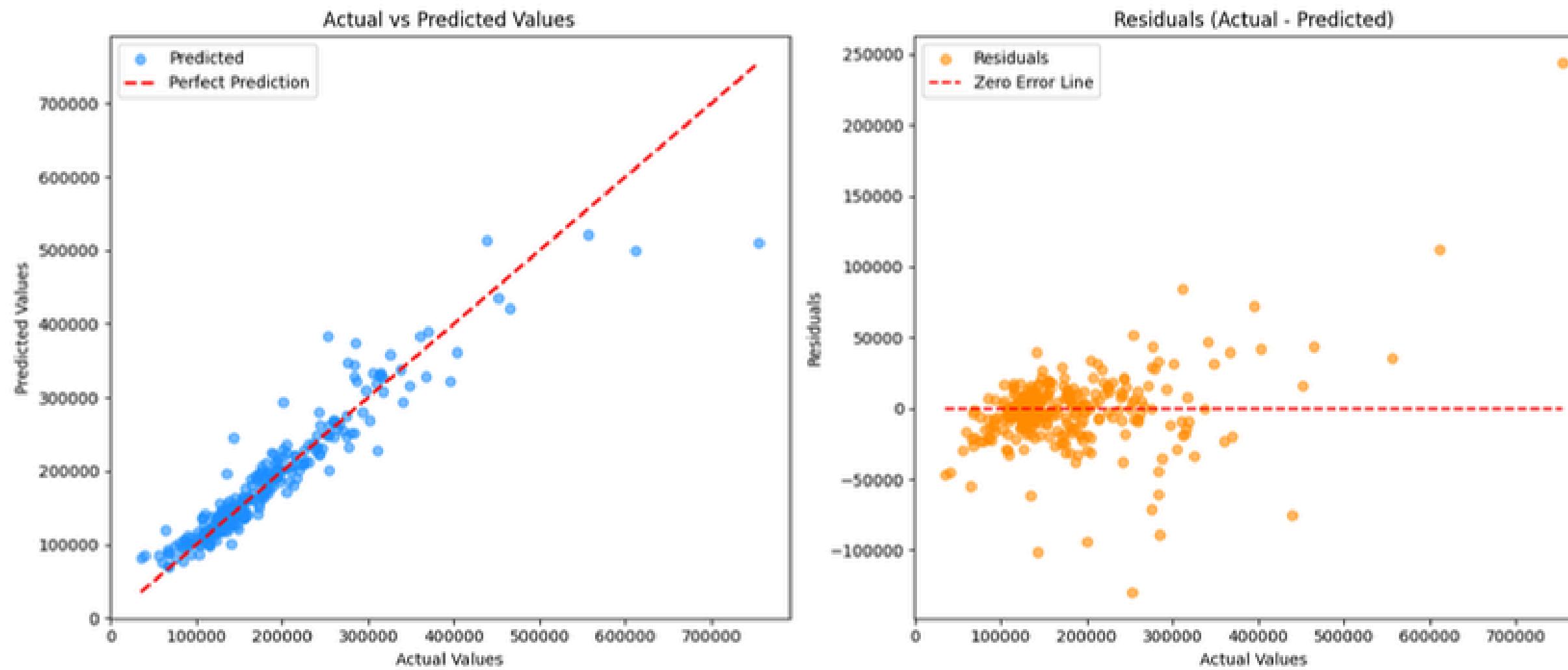


Results

Feature Importance



Results



- Plot sebelah kiri menunjukkan seberapa baik prediksi model mendekati garis prediksi sempurna (garis merah putus-putus). Titik-titik biru mendekati garis ini, menunjukkan bahwa model mampu memprediksi dengan cukup akurat, meskipun terdapat beberapa outlier yang jauh dari garis.
- Plot sebelah kanan menunjukkan distribusi residual (selisih antara nilai aktual dan prediksi). Sebagian besar residual berada di sekitar garis nol (garis merah), menunjukkan bias yang minim. Namun, terdapat beberapa titik residual yang besar, yang mengindikasikan prediksi yang kurang akurat untuk beberapa data.
- Kedua plot menunjukkan beberapa titik outlier dengan error yang signifikan, yang mungkin disebabkan oleh data yang tidak umum atau pola yang tidak tertangkap oleh model.

Real-world Application

The screenshot displays the rumah123 website's home loan simulation tool. Key details from the simulation are:

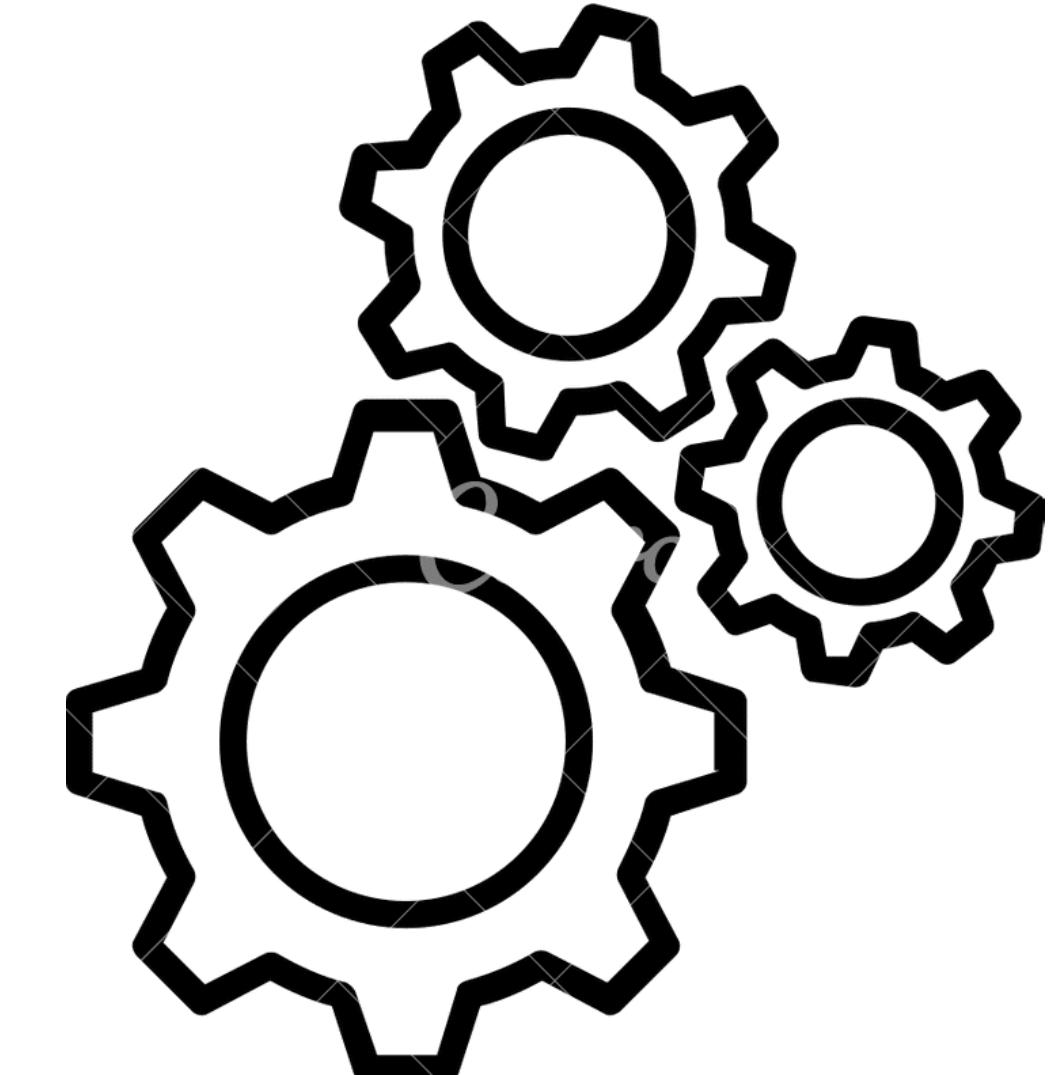
- Harga Properti: Rp 750,000,000
- Uang Muka: Rp 150,000,000 (20% of total price)
- Jangka Waktu KPR: 15 Tahun
- Pilihan Suku Bunga: BFI Finance | Bunga Fix: 7.92% | Masa Fix: 5 Tahun
- Angsuran/bulan Fix: Rp 5,706,237
- Tahun ke-1 Bunga: 7.92%

- Perkiraan harga rumah yang lebih akurat
 - Menggunakan Model machine learning , berdasarkan karakteristik properti, seperti lokasi, luas tanah, dan kondisi bangunan. Ini membantu pembeli membuat keputusan yang lebih tepat dan menghindari overpricing, serta membantu penjual menentukan harga jual yang kompetitif.
- Urban Development and Policy Planning
 - Perencanaan Kota dan Zonasi,
 - Program Perumahan Terjangkau
- Analisis Pasar Real Estate
 - Penilaian Properti,
 - Analisis Investasi
- Analisis Pinjaman Keuangan
 - Penilaian Risiko bagi Pemberi Pinjaman,
 - Opsi Refinance

Future Improvement

Solving current limitations of the provided solution.

- **Oversampling Data dengan Banyak Nilai Null:** Terapkan teknik oversampling pada data yang memiliki nilai Null tinggi untuk memastikan semua informasi tetap terwakili dalam model. Ini membantu mengurangi bias dan meningkatkan akurasi prediksi.
- **Klasifikasi Tambahan untuk Ketepatan Prediksi Harga:** Lakukan klasifikasi tambahan pada tipe rumah (misalnya, Premium, Medium, Standar) agar model dapat mempertimbangkan variasi harga berdasarkan kategori properti, menghasilkan estimasi yang lebih akurat.
- **Transformasi Logaritmik pada Sale Price:** Gunakan transformasi logaritmik pada variabel SalePrice untuk mengatasi ketimpangan distribusi data, yang membantu model lebih stabil dan presisi dalam prediksi.
- **Deployment Model untuk Aplikasi Real-Time:** Setelah pelatihan dan optimasi, integrasikan model ke dalam aplikasi real estate atau sistem penilaian harga rumah. Agen properti dan calon pembeli dapat mengakses estimasi harga secara instan, mempercepat pengambilan keputusan.
- **Penanganan Outliers pada Data Harga:** Identifikasi dan atasi outliers dalam data harga untuk mengurangi potensi distorsi dalam prediksi. Ini juga membantu model menggeneralisasi dengan lebih baik pada data baru.
- **Analisis Fitur Lingkungan (Lokasi, Fasilitas, dll.):** Libatkan informasi lingkungan seperti kualitas lokasi, fasilitas sekitar, dan akses transportasi. Fitur-fitur ini memberikan konteks tambahan pada harga properti dan menambah keakuratan prediksi.
- **Automasi Pembaruan Model dengan Data Baru:** Rancang pipeline untuk otomatisasi pembaruan model secara berkala dengan data terbaru, menjaga relevansi dan akurasi model terhadap tren pasar.



Conclusion

- Banyaknya variabel membuat penghapusan atau imputasi missing values rumit karena setiap variabel memerlukan perlakuan khusus tergantung tipe dan dampaknya.
- Ketidakseimbangan data di variabel MiscFeatures dan Pool menyebabkan model lebih fokus pada rumah standar, sehingga sulit menangkap pola untuk rumah premium atau kategori ekstrem lainnya.





Terima Kasih