# Dokumentacja Analizy Danych i Budowy Modelu Predykcyjnego

# Projektu: HouseWise

Autorzy: **Mikołaj Bernaszuk, Jacek Kazalski, Mikołaj Antoszewski, Tomek Małkiński, Mikołaj Kowaszewicz**

## Motywacja do stworzenia aplikacji

Główną motywacją do stworzenia aplikacji HouseWise była potrzeba uproszczenia procesu wyceny nieruchomości, który często bywa czasochłonny, skomplikowany i kosztowny. Dzięki wykorzystaniu uczenia maszynowego i nowoczesnych technologii, aplikacja ma na celu:

1. Demokratyzację dostępu do wiedzy o rynku nieruchomości
   Chcemy zapewnić każdemu użytkownikowi możliwość uzyskania rzetelnych prognoz cen nieruchomości bez konieczności angażowania specjalistów czy korzystania z drogich usług.
2. Wsparcie w podejmowaniu świadomych decyzji
   Nasza aplikacja pomaga użytkownikom lepiej zrozumieć, jakie cechy nieruchomości mają największy wpływ na jej wartość, umożliwiając lepsze planowanie i negocjacje.
3. Automatyzacja i oszczędność czasu
   Tradycyjne metody wyceny są czasochłonne i wymagają specjalistycznej wiedzy. HouseWise eliminuje te bariery, oferując szybkie i łatwe w użyciu narzędzie.
4. Zwiększenie dostępności technologii uczenia maszynowego
   Wprowadzenie modelu opartego na RandomForestRegressor pozwala wykorzystać zaawansowane algorytmy w sposób dostępny dla przeciętnego użytkownika, nawet bez specjalistycznej wiedzy technicznej.

Aplikacja została stworzona z myślą o tym, by uprościć proces wyceny, zwiększyć dostępność danych i pomóc użytkownikom podejmować bardziej świadome decyzje na dynamicznie zmieniającym się rynku nieruchomości.

# Spis treści

## Wstęp

Celem analizy było przeprowadzenie eksploracji danych dotyczących cen nieruchomości, identyfikacja kluczowych cech wpływających na cenę sprzedaży oraz stworzenie modelu predykcyjnego wykorzystującego algorytm Random Forest. W niniejszym dokumencie przedstawiono szczegółowe wyniki analizy danych, wizualizacje oraz interpretację wyników modelu.

## Opis Danych

Zbiór danych zawiera 1460 obserwacji i 11 zmiennych, z czego 8 to zmienne numeryczne, a 3 są kategoryczne. Kluczowe zmienne w zestawie danych to:

- OverallQual: Ogólna jakość wykończenia domu (skala 1-10).
- GarageCars: Liczba samochodów, które mogą być zaparkowane w garażu.
- ExterQual: Ocena jakości wykończenia zewnętrznego (kategoryczna).
- GrLivArea: Powierzchnia mieszkalna nadziemna (w stopach kwadratowych).
- FullBath: Liczba pełnych łazienek.
- KitchenQual: Jakość kuchni (kategoryczna).
- YearBuilt: Rok budowy nieruchomości.
- 1stFlrSF: Powierzchnia pierwszego piętra (w stopach kwadratowych).
- BsmtQual: Jakość piwnicy (kategoryczna).
- Fireplaces: Liczba kominków.
- SalePrice: Cena sprzedaży nieruchomości (zmienna docelowa).

- **Przykład danych:**

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | OverallQual | GarageCars | ExterQual | GrLivArea | FullBath | KitchenQual | YearBuilt | 1stFlrSF | BsmtQual | Fireplaces | SalePrice |
| 2 | 7 | 2 | Gd | 1710 | 2 | Gd | 2003 | 856 | Gd | 0 | 208500 |
| 3 | 6 | 2 | TA | 1262 | 2 | TA | 1976 | 1262 | Gd | 1 | 181500 |
| 4 | 7 | 2 | Gd | 1786 | 2 | Gd | 2001 | 920 | Gd | 1 | 223500 |
| 5 | 7 | 3 | TA | 1717 | 1 | Gd | 1915 | 961 | TA | 1 | 140000 |
| 6 | 8 | 3 | Gd | 2198 | 2 | Gd | 2000 | 1145 | Gd | 1 | 250000 |
| 7 | 5 | 2 | TA | 1362 | 1 | TA | 1993 | 796 | Gd | 0 | 143000 |
| 8 | 8 | 2 | Gd | 1694 | 2 | Gd | 2004 | 1694 | Ex | 1 | 307000 |
| 9 | 7 | 2 | TA | 2090 | 2 | TA | 1973 | 1107 | Gd | 2 | 200000 |
| 10 | 7 | 2 | TA | 1774 | 2 | TA | 1931 | 1022 | TA | 2 | 129900 |
| 11 | 5 | 1 | TA | 1077 | 1 | TA | 1939 | 1077 | TA | 2 | 118000 |
| 12 | 5 | 1 | TA | 1040 | 1 | TA | 1965 | 1040 | TA | 0 | 129500 |
| 13 | 9 | 3 | Ex | 2324 | 3 | Ex | 2005 | 1182 | Ex | 2 | 345000 |
| 14 | 5 | 1 | TA | 912 | 1 | TA | 1962 | 912 | TA | 0 | 144000 |
| 15 | 7 | 3 | Gd | 1494 | 2 | Gd | 2006 | 1494 | Gd | 1 | 279500 |
| 16 | 6 | 1 | TA | 1253 | 1 | TA | 1960 | 1253 | TA | 1 | 157000 |
| 17 | 7 | 2 | TA | 854 | 1 | TA | 1929 | 854 | TA | 0 | 132000 |
| 18 | 6 | 2 | TA | 1004 | 1 | TA | 1970 | 1004 | TA | 1 | 149000 |
| 19 | 4 | 2 | TA | 1296 | 2 | TA | 1967 | 1296 | | 0 | 90000 |
| 20 | 5 | 2 | TA | 1114 | 1 | Gd | 2004 | 1114 | TA | 0 | 159000 |
| 21 | 5 | 1 | TA | 1339 | 1 | TA | 1958 | 1339 | TA | 0 | 139000 |
| 22 | 8 | 3 | Gd | 2376 | 3 | Gd | 2005 | 1158 | Ex | 1 | 325300 |
| 23 | 7 | 1 | TA | 1108 | 1 | Gd | 1930 | 1108 | TA | 1 | 139400 |
| 24 | 8 | 2 | Gd | 1795 | 2 | Gd | 2002 | 1795 | Gd | 1 | 230000 |
| 25 | 5 | 2 | TA | 1060 | 1 | TA | 1976 | 1060 | Gd | 1 | 129900 |
| 26 | 5 | 1 | TA | 1060 | 1 | Gd | 1968 | 1060 | TA | 1 | 154000 |
| 27 | 8 | 3 | Gd | 1600 | 2 | Gd | 2007 | 1600 | Gd | 1 | 256300 |
| 28 | 5 | 2 | TA | 900 | 1 | Gd | 1951 | 900 | TA | 0 | 134800 |
| 29 | 8 | 3 | Gd | 1704 | 2 | Gd | 2007 | 1704 | Ex | 1 | 306000 |
| 30 | 5 | 1 | TA | 1600 | 1 | TA | 1957 | 1600 | TA | 2 | 207500 |
| 31 | 4 | 1 | TA | 520 | 1 | Fa | 1927 | 520 | TA | 0 | 68500 |
| 32 | 4 | 1 | TA | 1317 | 1 | TA | 1920 | 649 | TA | 0 | 40000 |
| 33 | 5 | 1 | TA | 1228 | 1 | Gd | 1966 | 1228 | TA | 0 | 149350 |
| 34 | 8 | 2 | Gd | 1234 | 2 | Gd | 2007 | 1234 | Ex | 0 | 179900 |
| 35 | 5 | 2 | TA | 1700 | 1 | Gd | 1959 | 1700 | TA | 1 | 165500 |
| 36 | 9 | 2 | Ex | 1561 | 2 | Ex | 2005 | 1561 | Ex | 1 | 277500 |
| 37 | 8 | 3 | Gd | 2452 | 3 | Gd | 2004 | 1132 | Ex | 1 | 309000 |

## Eksploracyjna Analiza Danych

### Analiza wartości odstających

Box ploty dla zmiennych numerycznych ujawniły obecność wartości odstających w takich zmiennych jak GrLivArea, 1stFlrSF i SalePrice. Wartości te mogą mieć wpływ na modelowanie, jednak nie zostały usunięte, aby zachować pełne spektrum danych.

### Macierz korelacji

Analiza korelacji ujawniła, że zmienna SalePrice wykazuje najsilniejszą dodatnią korelację z:

- OverallQual (0.79)
- GrLivArea (0.71)
- 1stFlrSF (0.61)

## Modelowanie

### Przygotowanie danych

- Proces oczyszczania danych rozpoczął się od uzupełnienia brakujących wartości z wykorzystaniem klasy SimpleImputer(). W przypadku danych numerycznych zastosowano strategię uzupełniania braków za pomocą średniej, natomiast dla danych kategorycznych użyto strategii zastępowania braków najczęściej występującą wartością. Kolejnym krokiem była normalizacja danych numerycznych przy użyciu StandardScaler(), która ujednoliciła ich skalę, a także zakodowanie danych kategorycznych za pomocą OneHotEncoder, co przekształciło je w wektory binarne. W celu zintegrowania tych działań dla różnych typów danych wykorzystano ColumnTransformer, pozwalając na równoczesne przetwarzanie cech numerycznych i kategorycznych w jednym procesie.
- Po wstępnym przetworzeniu danych oszacowano ważność cech z wykorzystaniem modelu regresji lasów losowych. Na tej podstawie wybrano 10 najistotniejszych zmiennych, które miały największy wpływ na jakość predykcji, a pozostałe cechy zostały usunięte.
- Ostatecznie oczyszczony zbiór danych, zawierający wybrane cechy oraz zmienną docelową, został zapisany w pliku CSV jako cleaned_dataset.csv, stanowiąc gotowy materiał do dalszych analiz i modelowania.

## Narzędzie AutoML

W projekcie został użyty TPOT (Tree-based Pipeline Optimization Tool) – jest to narzędzie AutoML, które automatyzuje proces projektowania, optymalizacji i wybiera najlepsze modele, co pozwala na znaczną oszczędność czasu i zasobów podczas eksploracji różnych konfiguracji modeli. Dzięki TPOT użytkownicy mogą skupić się na analizie wyników, zamiast ręcznie dostrajać hiperparametry czy budować złożone potoki przetwarzania danych.

### Wyniki TPOT

Generation 1 - Current best internal CV score: -1092468711.620651

Generation 2 - Current best internal CV score: -1092468711.620651

Generation 3 - Current best internal CV score: -1084372773.9958704

Generation 4 - Current best internal CV score: -1023457211.0313208

Generation 5 - Current best internal CV score: -1023457211.0313208

#### Best pipeline

RandomForestRegressor(StandardScaler(input_matrix), bootstrap=False, max_features=0.25, min_samples_leaf=1, min_samples_split=7, n_estimators=100))

## Model

- Model predykcyjny został stworzony za pomocą algorytmu RandomForestRegression z parametami wybranymi przez narzędzie automl TpotRegressor:
  a) bootstrap=False,
  b) max_features=0.25,
  c) min_samples_leaf=1,
  d) min_samples_split=7,
  e) n_estimators=100.
- Najważniejsze cechy wpływające na predykcję ceny:
  1. OverallQual: 59.66%
  2. GrLivArea: 16.82%
  3. 1stFlrSF: 9.02%
  4. YearBuilt: 4.96%
  5. GarageCars: 3.64%

  Największy wpływ na cenę sprzedaży ma ogólna jakość nieruchomości (OverallQual)

## Wyniki Modelu:

**MAE**: 19152.64

**MSE**: 889314952.12

**RMSE**: 29821.38

**R2**: 0.88405

**Interpretacja wyników:**

- Model osiągnął wysoki współczynnik determinacji $R^2$ (0.88405), co wskazuje, że dobrze wyjaśnia zmienność w cenach nieruchomości.
- Stosunkowo niski MAE (19,152.64) i RMSE (29,821.38) sugerują, że model działa precyzyjnie i większość przewidywanych cen znajduje się blisko wartości rzeczywistych.
- Model predykcyjny oparty na Random Forest dobrze sprawdza się w zadaniu przewidywania cen nieruchomości. Wysoki $R^2$ oraz niskie MAE i RMSE świadczą o jego skuteczności.

## Raport Profilowania Danych

Za pomocą narzędzia ydata_profiling wygenerowano szczegółowy raport zawierający:

- Podsumowanie statystyk.
- Identyfikację braków danych.
- Analizę dystrybucji zmiennych.

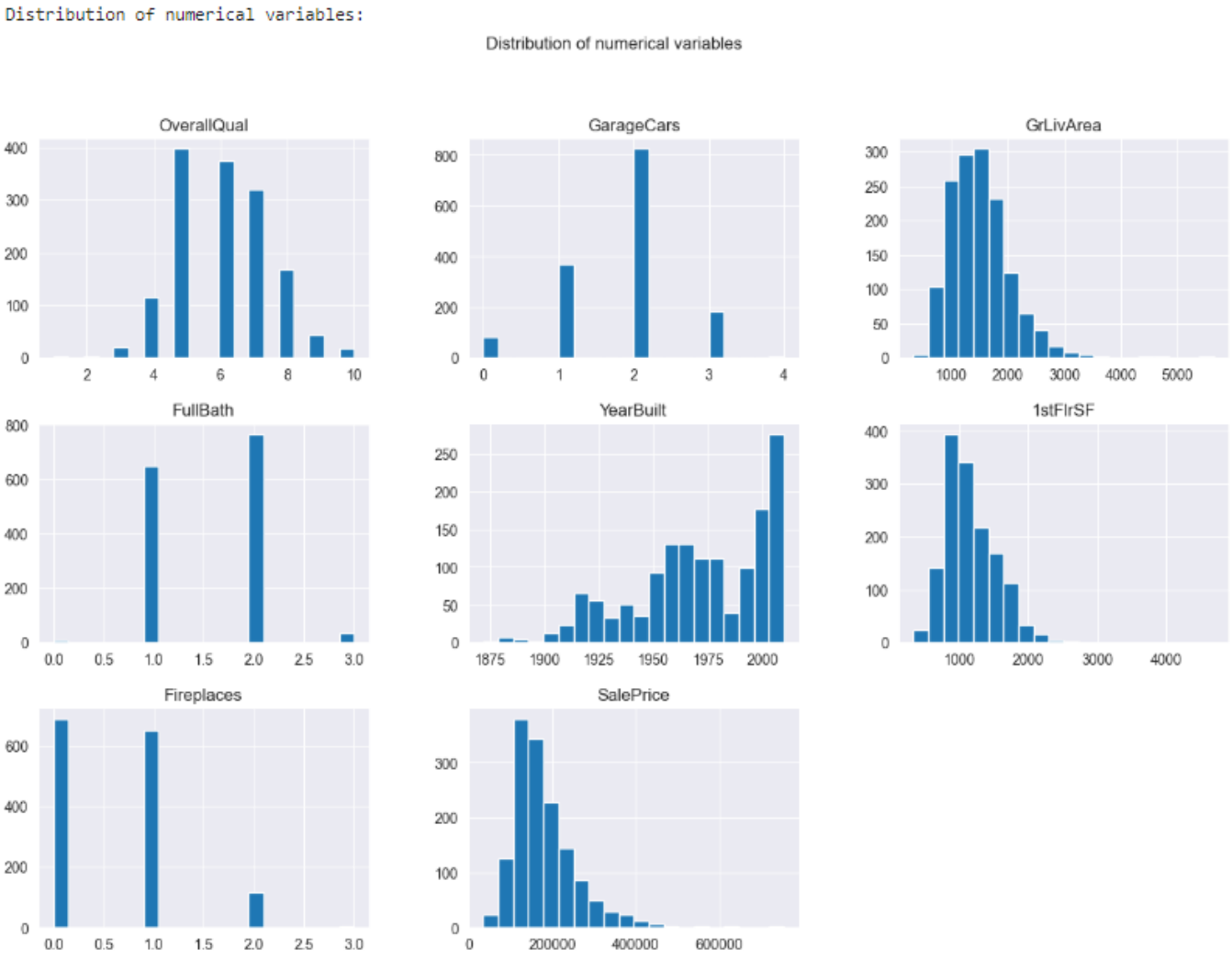
- Zależności między zmiennymi.

## Wnioski

1. Ogólna jakość wykończenia (OverallQual) oraz powierzchnia mieszkalna (GrLivArea) są kluczowymi czynnikami wpływającymi na cenę nieruchomości.
2. Zmienna SalePrice wykazuje asymetryczny rozkład, co może wpłynąć na dokładność predykcji.
3. Model Random Forest zidentyfikował cechy o największym wpływie na wynik, co może być wykorzystane do podejmowania decyzji biznesowych.
4. Wygenerowany raport profilowania danych dostarcza szczegółowych informacji do dalszych analiz.

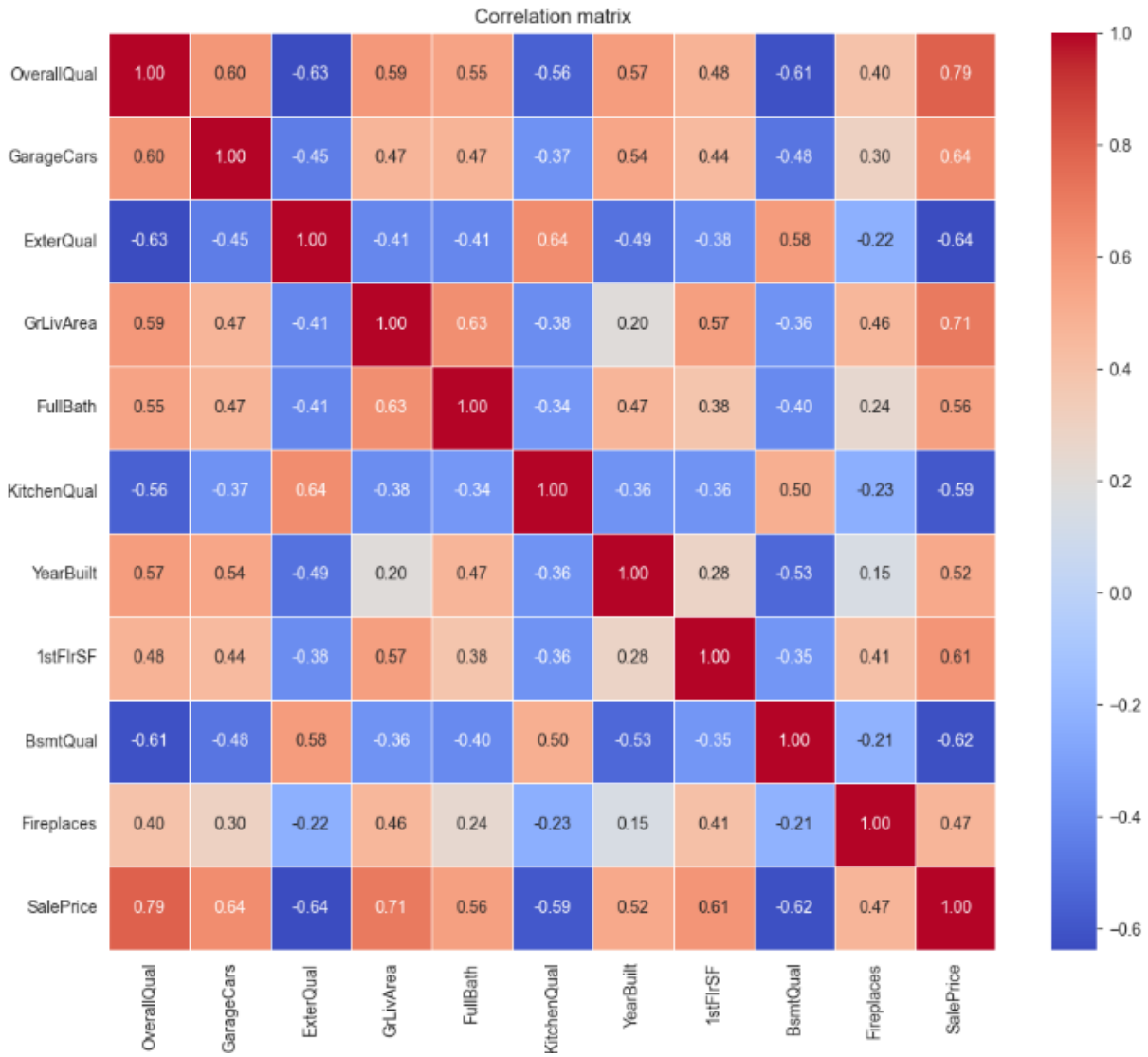
# Wykresy

## Wykres "Rozkładu zmiennych numerycznych”:



## Macierz korelacji:

Correlation matrix:



## Analiza Kodu w Projekcie

Ocena ogólna kodu:

- Użyliśmy do oceny kodu biblioteki zwanej pylint. Kod uzyskał ocenę 8.93/10. Widoczne błędy zostały pominięte, ponieważ dwa pierwsze wynikają z niepoprawnego rozpoznania importów przez pylint, a trzeci jest efektem świadomego wyboru projektowego dotyczącego struktury komunikacji między frontendem a backendem.

```
(HouseWise) PS C:\Users\mkowa\PycharmProjects\HouseWise> pylint . --ignore=frontend
************* Module backend.fastApiProject.main
backend\fastApiProject\main.py:12:0: E0401: Unable to import 'schemas.schema' (import-error)
backend\fastApiProject\main.py:13:0: E0401: Unable to import 'mappings.mapping' (import-error)
backend\fastApiProject\main.py:77:11: W0718: Catching too general exception Exception (broad-exception-caught)

-------------------------------------------------------------------
Your code has been rated at 8.93/10 (previous run: 8.93/10, +0.00)
```

## Dobór technologii

Czym jest Svelte użyte w projekcie?

- Svelte to nowoczesny framework JavaScript, który służy do tworzenia interfejsów użytkownika (UI). W odróżnieniu od popularnych frameworków takich jak React czy Vue, Svelte działa na etapie kompilacji, a nie w przeglądarce

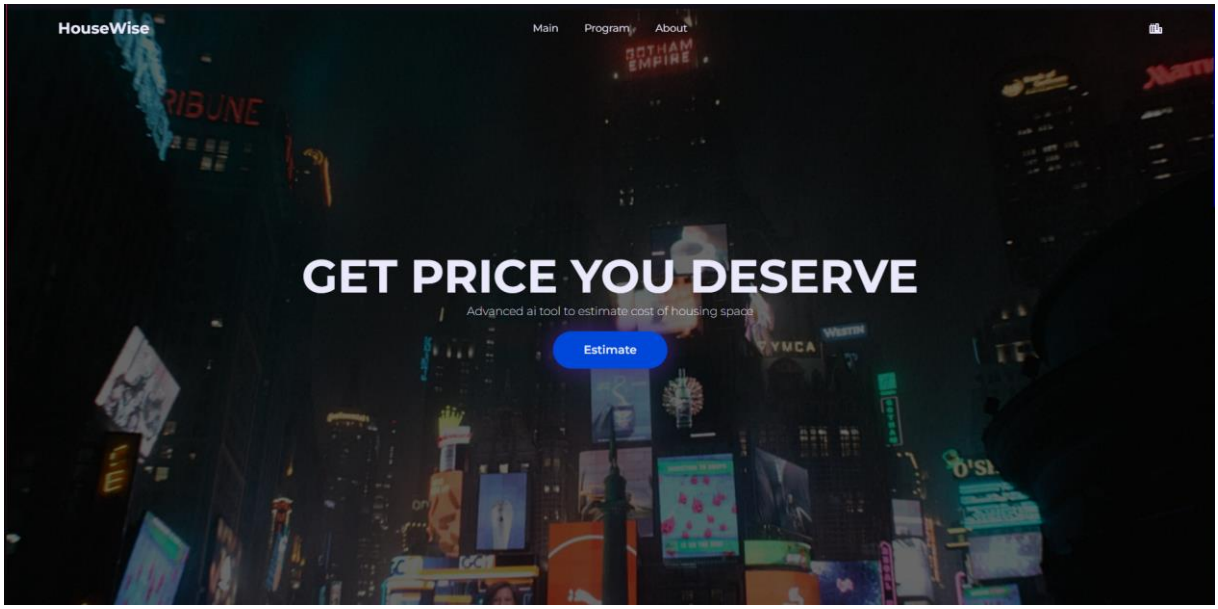
Czym jest FastApi użyte w projekcie?

- FastAPI to nowoczesny framework webowy dla Pythona, zaprojektowany z myślą o tworzeniu szybkich i wydajnych interfejsów API (REST i GraphQL). Wykorzystuje asynchroniczne funkcje Pythona (asyncio) oraz automatyczne generowanie dokumentacji.

## Widoki frontendu:

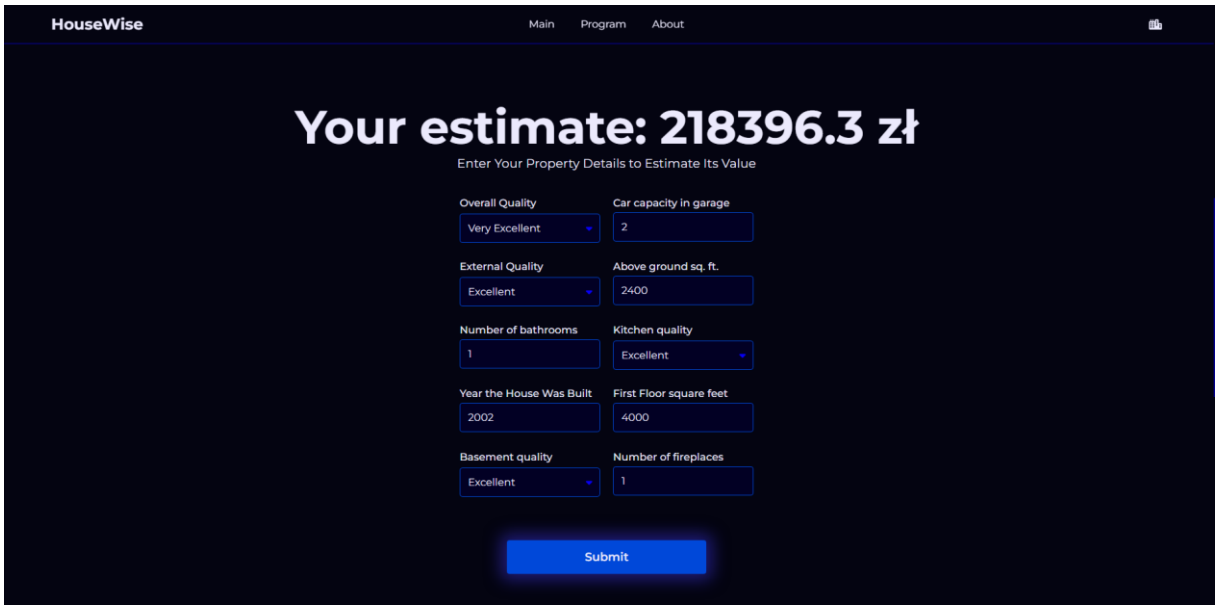
**Landing section:**

- Intuicyjny i łatwy w obsłudze widok, który pozwala użytkownikowi szybko przejść do przewidywania ceny nieruchomości. Całość uzupełnia chwytliwy slogan przyciągający uwagę i jasno wyjaśniający cel aplikacji.



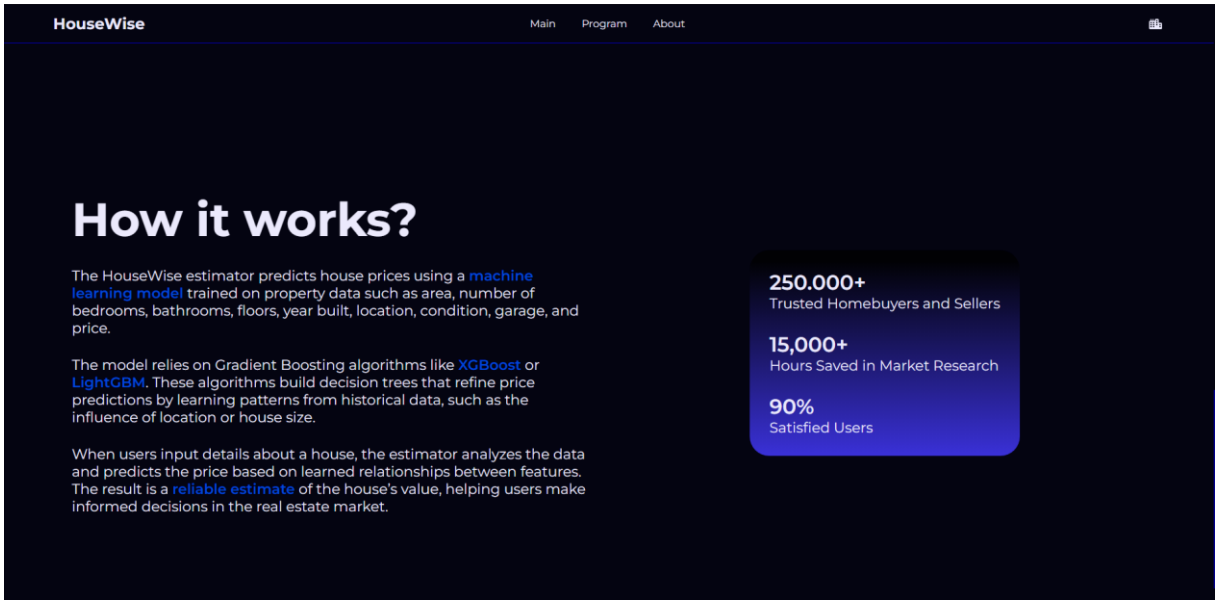

**Estimator section:**

- Estymator to centralny element aplikacji, będący formularzem zaprojektowanym z myślą o użytkowniku. Oferuje pełną walidację danych, wyświetlanie komunikatów o błędach w czasie rzeczywistym oraz intuicyjny design.



**About section:**

- About Section opisuje sposób działania aplikacji HouseWise. Estymator przewiduje ceny nieruchomości na podstawie modelu uczenia maszynowego opartego na algorytmie RandomForestRegressor. Model wykorzystuje dane o nieruchomości, takie jak pojemność garażu, ogólna jakość, liczba łazienek, rok budowy, powierzchnia mieszkalna, jakość wykończenia zewnętrznego, liczba kominków oraz jakość piwnicy i kuchni.




## Podsumowanie

Projekt **HouseWise** stanowi innowacyjne podejście do wyceny nieruchomości, łącząc zaawansowane technologie uczenia maszynowego z intuicyjnym interfejsem użytkownika. Dzięki zastosowaniu algorytmu RandomForestRegressor i szczegółowej analizie danych, aplikacja oferuje precyzyjne prognozy cen, które mogą wspierać użytkowników w podejmowaniu świadomych decyzji na rynku nieruchomości.

W trakcie realizacji projektu:

- Przeprowadzono dokładną eksplorację i analizę danych, identyfikując kluczowe cechy wpływające na cenę nieruchomości.
- Zbudowano model predykcyjny, który osiągnął wysoki poziom dokładności ($R^2$: 0.8916).
- Stworzono interfejs, który umożliwia łatwe wprowadzanie danych i szybkie uzyskanie prognoz.

Dzięki połączeniu najnowszych technologii, takich jak **Svelte** i **FastAPI**, z solidnym fundamentem analizy danych i modelowania, aplikacja HouseWise stanowi przykład, jak nowoczesne narzędzia mogą wspierać realne potrzeby użytkowników.