Predicting California Housing Prices Using Ensemble Models

Project Agenda

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion

Project Agenda - Part 1

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion

1. Executive Summary

- Background
- Related Works
- Data Relationships

Data Understanding

- Numeric Data
- Category Data
- Time and Location

Data Preprocessing

- Voting
- Boosting
- Stacking

Data Modeling

- Model Comparision
- Model Enhancement

Data Evaluation

Project Agenda - Part 2

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion

2. Project Introduction

Problem Statement	Related Works	Project Timeline
✓ Background	✓ Leaderboard	✓ Proposal
✓ Data Source	✓ Performance	✓ Checkpoint
✓ Objectives	✓ Methods	✓ Final

2.1 Problem Statement

Background

- Kaggle Competition Hosted by D2L
- Predicting California House Prices
- Evaluated by RMSE of Log Predictions

Data Source

- Real-world Data Scraped from Zillow.com
- Including Errors and Noises
- Training 2020 all Data to Predict 2021

Objectives

- Extracting the Feature Correlations
- Improving Model Performance
- Surpassing the #1 Score

2.2 Related Works

AutoML and Ensemble models perform best on this dataset.

#	△	Team		Score
1	—	fxzero		0.11922
2	▲ 1	sxjscience	AutoML - AutoGluon	0.12063
3	▲ 1	wuwawa	AutoML - h20	0.12283
4	▲ 5	Leon		0.12283
5	▲ 6	jack	Random Forest	0.12455
6	▲ 2	totoro		0.12485
7	▲ 8	automl (baseline)	AutoML - AutoGluon	0.12502
8	▼ 2	pullpullyes		0.12518
9	▲ 5	Shaoqing		0.12559
10	—	haoxinkuhaoxiangku		0.12583
11	▼ 4	pppp		0.12588
12	▲ 1	TonightEatChicken	CatBoost	0.12632

# of Teams: 173

 # 1 Score: 0.11922
 # 18 Score: 0.13003
 # 30 Score: 0.14060
 # 39 Score: 0.15097
 # 48 Score: 0.16211

Best Models:
Auto ML Models
Ensemble Modles

2.3 Project Timeline

Improving model accuracy and Surpassing #1 performance

Proposal:
- Project Structure
- Model Selection
- Task Objectives

Checkpoint:
- Process Tracking
- Model Evaluation
- Report Writing

Final:
- Project Summary
- Future Enhancement
- Report Completed

1st Week
Top20%/<0.15

⇨

2nd Week
Top10%/<0.13

⇨

3rd Week
Top1%/<0.11

Project Agenda - Part 3

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion

3. Data Preprocessing

☐ Data Cleaning and Transformation impact the performance

#	Column	Non-Null Count	Dtype
0	Id	79065 non-null	int64
1	Address	79065 non-null	object
2	Sold Price	47439 non-null	float64
3	Summary	78226 non-null	object
4	Type	79065 non-null	object
5	Year built	77123 non-null	float64
6	Heating	67552 non-null	object
7	Cooling	63956 non-null	object
8	Parking	77389 non-null	object
9	Lot	56076 non-null	float64
10	Bedrooms	74467 non-null	object
11	Bathrooms	73655 non-null	float64
12	Full bathrooms	66137 non-null	float64
13	Total interior livable area	75187 non-null	float64
14	Total spaces	77398 non-null	float64
15	Garage spaces	77398 non-null	float64
16	Region	79063 non-null	object
17	Elementary School	70572 non-null	object
18	Elementary School Score	70330 non-null	float64
19	Elementary School Distance	70572 non-null	float64
20	Middle School	50788 non-null	object
21	Middle School Score	50786 non-null	float64
22	Middle School Distance	50788 non-null	float64
23	High School	71891 non-null	object
24	High School Score	71281 non-null	float64
25	High School Distance	71890 non-null	float64
26	Flooring	57138 non-null	object
27	Heating features	66517 non-null	object
28	Cooling features	62432 non-null	object
29	Appliances included	55716 non-null	object
30	Laundry features	59083 non-null	object
31	Parking features	72437 non-null	object
32	Tax assessed value	72742 non-null	float64
33	Annual tax amount	71856 non-null	float64
34	Listed On	79065 non-null	object
35	Listed Price	79065 non-null	float64
36	Last Sold On	49520 non-null	object
37	Last Sold Price	49520 non-null	float64
38	City	79065 non-null	object
39	Zip	79065 non-null	int64
40	State	79065 non-null	object

40 Features

Train : 47439 Samples
Test : 31626 Samples

Numeric Features

Price: List, Last Sold, Tax
Number: Bedr, Bathr, Space

Category Features

Type, Heating, Cooling,
Parking, Flooring, Laundry

Time & Location

Time: Build, List, Last Sold
Location: Region, City, Zip

3.1 Numerical Features

Predicting by Listed Price and Optimizing by other features

Correlations

	Correlation	Mean
Sold Price	1.0000	1175553.3
Listed Price	0.8128	1207590.9
Last Sold Price	0.6284	736396.0
Annual tax amount	0.5545	8959.9
Tax assessed value	0.5502	702650.4
Bathrooms	0.3916	2.3
Full bathrooms	0.3750	2.0
Bedrooms	0.3008	2.9
Elementary School Score	0.2756	5.7
Total interior livable area	0.2486	1846.7
Middle School Score	0.2443	5.3
High School Score	0.1916	6.1
Garage spaces	0.0560	1.4
Total spaces	0.0393	1.4
Lot	-0.0106	51034.9
Middle School Distance	-0.0110	1.4
Year built	-0.0155	1966.1
Elementary School Distance	-0.0626	0.9
High School Distance	-0.0684	2.0

Removing Outliers

	Correlation	Mean
Sold Price	1.0000	1296050.5
Annual tax amount	0.7462	9956.8
Tax assessed value	0.7432	786311.8
Last Sold Price	0.7068	807853.7
Listed Price	0.6127	1315890.3
Full bathrooms	0.5442	2.1
Bathrooms	0.5162	2.4
Bedrooms	0.3069	3.0
Elementary School Score	0.2756	5.7
Middle School Score	0.2443	5.3
High School Score	0.1916	6.1
Garage spaces	0.0093	1.5
Total spaces	0.0075	1.6
Total interior livable area	-0.0015	5774.6
Lot	-0.0066	235338.3
Year built	-0.0271	1956.6
Middle School Distance	-0.0593	1.7
Elementary School Distance	-0.0862	1.2
High School Distance	-0.0995	2.4

With Outliers

Key Features:

- Price Related Features

Support Features:

- Most Numeric Features

Trade-off:

- Dealing with Extreme Values

Enhancement:

- Filling Missing Values

3.2 Category Features

Transform category features considering bias and variance

Before Reduction	After Reduction
Type: 174 unique values	Type: 7 unique values
Flooring: 1739 unique values	Flooring: 10 unique values
Heating features: 1762 unique values	Heating features: 16 unique values
Cooling features: 595 unique values	Cooling features: 8 unique values
Laundry features: 3030 unique values	Laundry features: 13 unique values
Parking features: 9694 unique values	Parking features: 18 unique values
Appliances included: 11289 unique values	Appliances included: 10 unique values

Cause:	Scraping Method	Website Structure	Input Errors
Solution:	Domain Knowledge	Balance Distribution	Monitor Variance

3.3 Data Transformaton

- Transform the data for futher modeling

Time data
Aggregation by Year

Location data
Aggregation by Zip(3)

Category data
One-hot Dummy Variable

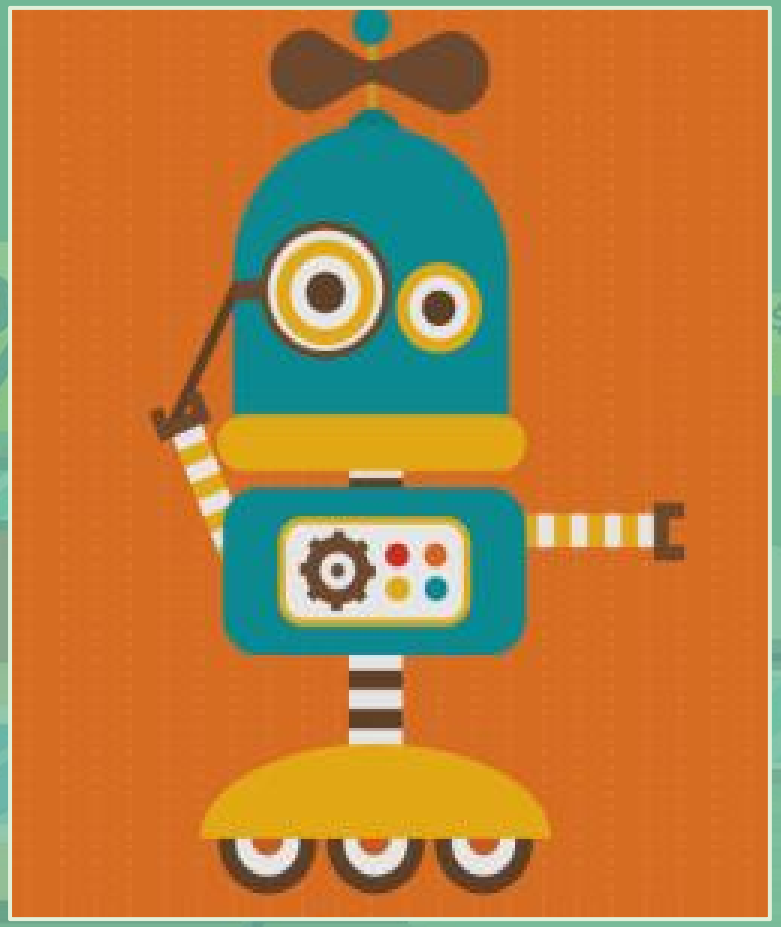
Numerical data
Log Transformation

Extreme data
Replace by Nan

Missing data
Replace by Nan

Project Agenda - Part 4

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion



4. Model Evaluation

- Using Listed Price as Prediction, the score achieves #38

Guessing

Train Log_RMSE:
0.11511
Val Log_RMSE:
0.12323
Test Log_RMSE:

Voting

Random Forest

Boosting

CatBoost

Stacking

Combine Models

4.1 Voting - Random Forest

After tuning hyper parameters, the score achieves #12

Random Forest (Default)

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': None,
 'max_features': 1.0,
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 26,
 'verbose': 0,
 'warm_start': False}
```

Train Log_RMSE:
0.03578
Val Log_RMSE:
0.11736
Test Log_RMSE:

RF-Default.csv 0.13771
Complete (after de...

Significant Overftiting

Random Forest (Tuning)

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': 18,
 'max_features': 1.0,
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 9,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 600,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 26,
 'verbose': 0,
 'warm_start': False}
```

Train Log_RMSE:
0.08056
Val Log_RMSE:
0.11420
Test Log_RMSE:

RF-Tuning.csv 0.12608
Complete (after de...

Slight Overfitting

4.2 Boosting - CatBoost

After tuning hyper parameters, the score achieves #6

CatBoost (Default)

```
'iterations': 1000,
'sampling_frequency': 'PerTree',
'leaf_estimation_method': 'Newton',
'random_score_type': 'NormalWithModelSizeDecrease',
'grow_policy': 'SymmetricTree',
'penalties_coefficient': 1,
'boosting_type': 'Plain',
'model_shrink_mode': 'Constant',
'feature_border_type': 'GreedyLogSum',
'bayesian_matrix_reg': 0.10000000149011612,
'eval_fraction': 0,
'force_unit_auto_pair_weights': False,
'l2_leaf_reg': 3,
'random_strength': 1,
'rsm': 1,
'boost_from_average': True,
'model_size_reg': 0.5,
'pool_metainfo_options': {'tags': {}},
'subsample': 0.800000011920929,
'use_best_model': False,
'random_seed': 26,
'depth': 6,
'posterior_sampling': False,
'border_count': 254,
'classes_count': 0,
'auto_class_weights': 'None',
'sparse_features_conflict_fraction': 0,
'leaf_estimation_backtracking': 'AnyImprovement',
'best_model_min_trees': 1,
'model_shrink_rate': 0,
'min_data_in_leaf': 1,
'loss_function': 'RMSE',
'learning_rate': 0.07409299910068512,
```

Train Log_RMSE:
0.08235
Val Log_RMSE:
0.10884
Test Log_RMSE:

CB-Default.csv 0.12751
Complete (after de...

SlightOverfitting

CatBoost (Tuning)

```
'iterations': 1700,
'sampling_frequency': 'PerTree',
'leaf_estimation_method': 'Newton',
'random_score_type': 'NormalWithModelSizeDecrease',
'grow_policy': 'SymmetricTree',
'penalties_coefficient': 1,
'boosting_type': 'Plain',
'model_shrink_mode': 'Constant',
'feature_border_type': 'GreedyLogSum',
'bayesian_matrix_reg': 0.10000000149011612,
'eval_fraction': 0,
'force_unit_auto_pair_weights': False,
'l2_leaf_reg': 3,
'random_strength': 1,
'rsm': 1,
'boost_from_average': True,
'model_size_reg': 0.5,
'pool_metainfo_options': {'tags': {}},
'subsample': 0.800000011920929,
'use_best_model': False,
'random_seed': 26,
'depth': 10,
'posterior_sampling': False,
'border_count': 254,
'classes_count': 0,
'auto_class_weights': 'None',
'sparse_features_conflict_fraction': 0,
'leaf_estimation_backtracking': 'AnyImprovement',
'best_model_min_trees': 1,
'model_shrink_rate': 0,
'min_data_in_leaf': 1,
'loss_function': 'RMSE',
'learning_rate': 0.023000000044703484,
```

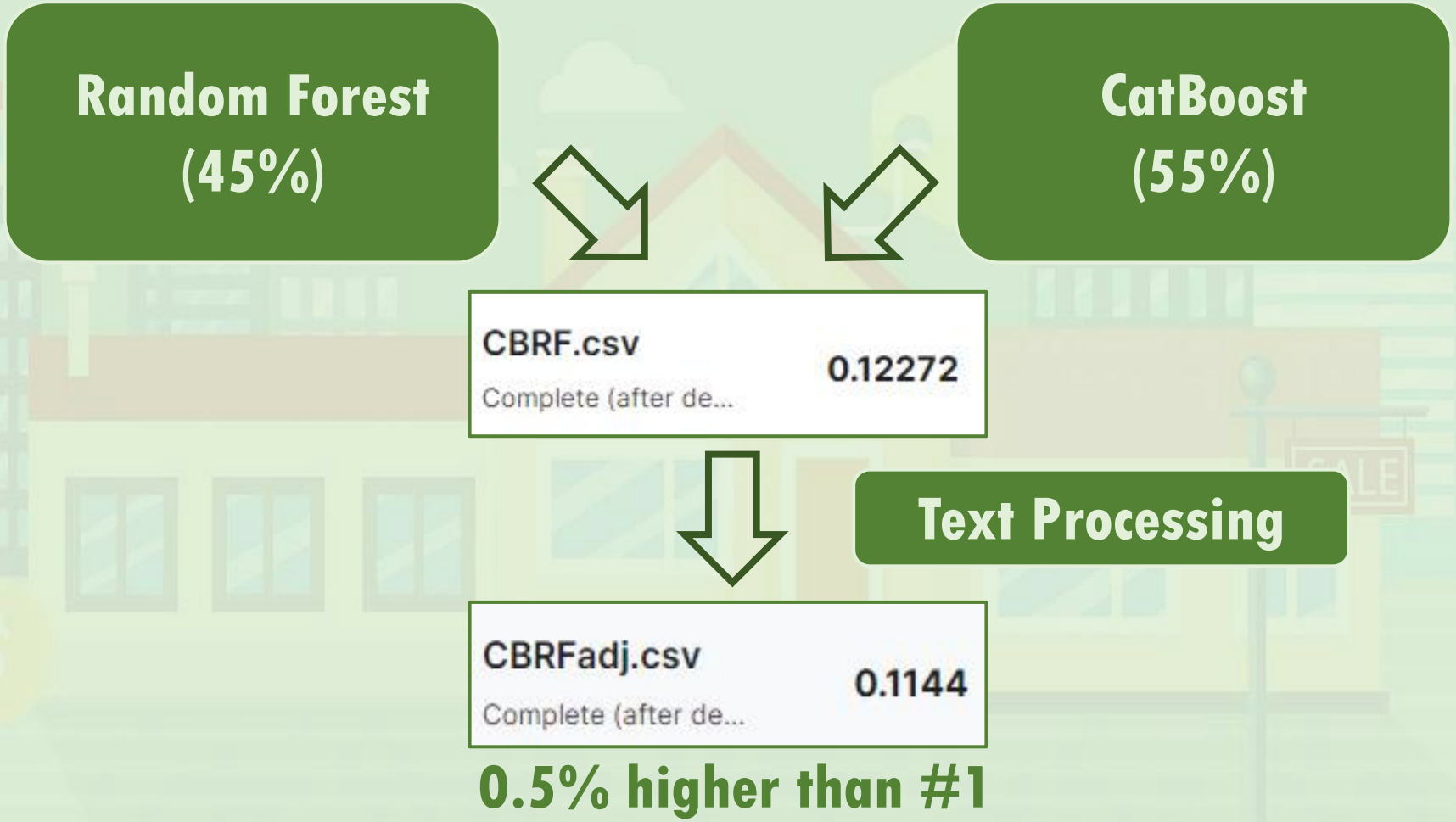
Train Log_RMSE:
0.06969
Val Log_RMSE:
0.10514
Test Log_RMSE:

CB-Tuning.csv 0.12458
Complete (after de...

SlightOverfitting

4.3 Stacking

- Combine two Models, the score achieves #3



Project Agenda - Part 5

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion

5. Future Enhancement

- Random Forest is a voting method (tree has equal weight), hence, important features become pronounced due to aggregation.

Random Forest

Feature	Importance
Listed Price	94.38
Tax assessed value	2.28
Listed On	0.68
Annual tax amount	0.44
Last Sold Price	0.28
Full bathrooms	0.25
Bathrooms	0.23
Year built	0.15
Total interior livable area	0.13
High School Distance	0.11
Lot	0.10
Last Sold On	0.10
Elementary School Distance	0.08
Elementary School Score	0.08
Zip_900	0.08
Middle School Distance	0.07
High School Score	0.06
Appliances included	0.06
Parking features_GGAttach	0.05
Bedrooms	0.04

CatBoost

Feature	Importance
Listed Price	62.24
Listed On	7.03
Tax assessed value	4.34
Annual tax amount	3.93
Last Sold Price	2.02
Last Sold On	1.79
Year built	1.62
Bathrooms	1.40
Total interior livable area	1.19
Zip_900	1.06
High School Distance	0.92
Full bathrooms	0.92
Zip_951	0.88
Elementary School Distance	0.88
Lot	0.77
Middle School Distance	0.76
High School Score	0.72
Elementary School Score	0.70
Middle School Score	0.61
Appliances included	0.60

- CatBoost is a Boosting method (tree has varying weights), resulting in relatively increased importance for other features.

5. Future Enhancement

Validation Selection	No random selection Matching the Test Dataset
Feature Engineering	Create New Features: Room Price, Sq.feet Price
Extreme Values	Using Ensemle models to detect Extreme Value
Missing Values	Using Ensemle models to fill missing Value

Project Agenda - Part 6

1. Executive Summary
2. Project Introduction
3. Data Preprocessing
4. Model Evaluation
5. Future Enhancement
6. Project Conclusion

6. Project Conclusion

Data Understanding:
determines whether the path is correct

Data Preprocessing:
determines the accuracy of the outcome

Ensemble is Powerful:
but AutoML will be the future

Predicting California Housing Prices Using Ensemble Models