Predictive Analytics in Real Estate: Leveraging Housing Data to Forecast Market Trends and Revitalization Opportunities

---

By Ashish Mangal

Abstract

This study aims to explore the use of the "Housing & Ocean Proximity" dataset to predict key real estate market indicators. It focuses on forecasting housing prices, crime rates, and identifying areas for potential urban revamping. By integrating machine learning algorithms with comprehensive housing data, including neighborhood demographics and economic conditions, the study aims to offer actionable insights for investors, urban planners, and policy makers, contributing to a more responsive and informed approach in the dynamic real estate sector.

Keywords

Real Estate Analytics, Machine Learning, Housing Market trends, Urban Planning

Introduction

The rapid evolution of the real estate market, influenced by changing demographics, economic factors, and urbanization, necessitates advanced analytical approaches. This paper delves into the potential of machine learning in transforming real estate analytics. Utilizing the "Housing & Ocean Proximity" dataset, it aims to forecast housing prices, analyze crime rates, and identify areas ripe for revamping. The study underscores the importance of integrating data on house characteristics and neighborhood dynamics to predict market trends. By harnessing predictive modeling, the research seeks to offer valuable insights for stakeholders in real estate, urban planning, and policy-making, highlighting the synergy between data science and real estate market analysis.

Related Work:

Research Paper Citation

Paper 1: **Endogenous Gentrification and Housing Price Dynamics**

The paper discusses the variation in house price growth in different neighborhoods within a city during overall housing booms. It introduces a model linking these price movements and neighborhood gentrification following citywide housing demand shocks. A key concept is the preference for living near wealthier neighbors, leading to income-based segregation. Higher-income residents move into adjacent poorer areas during demand spikes, driving up prices and displacing original residents, a process termed "endogenous gentrification." The paper provides empirical evidence supporting this model, using various data sets and city-level demand shock analysis.

Paper 2: **Stratifying and predicting patterns of neighborhood change and gentrification: An urban analytics approach**

This paper tackles the complexity of identifying and differentiating gentrification from other types of neighborhood changes in cities, using London as a case study. It employs a novel urban analytics approach, integrating diverse datasets on population, house prices, and development. The study uses data reduction and classification methods, followed by machine learning, to analyze and predict gentrification trends.

Data Definitions

It is a dataset that contains information about houses and their proximity to the ocean. The dataset contains the following data features:

- **Address:** The address of the house.
- **Latitude and longitude:** The latitude and longitude of the house.
- **Distance to ocean:** The distance from the house to the ocean.
- **Type of house:** The type of house (e.g., single-family home, apartment, condo).
- **Number of bedrooms:** The number of bedrooms in the house.
- **Number of bathrooms:** The number of bathrooms in the house.
- **Square footage:** The square footage of the house.
- **Price:** The price of the house.

The data comes from a variety of sources, such as public records, real estate listings, and surveys.

Sources of housing ocean proximity data that I have got access to use, is from my previous CMU Project for an HCII elective. It is varied and includes coastal survey databases managed by governmental agencies, which provide detailed information about properties near the ocean. Real estate websites often feature filters and maps highlighting oceanfront or ocean-close properties. GIS data from national and local sources can offer precise mapping of property locations in relation to the coastline.

Baseline Performance

=== Run information ===

```
Scheme:       weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"
Relation:     h_700-weka.filters.unsupervised.attribute.Remove-R11
Instances:    700
Attributes:   10
              longitude
              latitude
              housing_median_age
              total_rooms
              total_bedrooms
              population
              households
              median_income
              median_house_value
              ocean_proximity
Test mode:    10-fold cross-validation
```

=== Classifier model (full training set) ===

```
SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: <1H OCEAN, INLAND

BinarySMO

Machine linear: showing attribute weights, not support vectors.

         4.8876 * (normalized) longitude
 +       6.3828 * (normalized) latitude
 +      -1.0439 * (normalized) housing_median_age
 +       2.4531 * (normalized) total_rooms
 +       0.3129 * (normalized) total_bedrooms
 +      -1.4328 * (normalized) population
 +      -0.5887 * (normalized) households
 +       1.1885 * (normalized) median_income
 +      -4.9497 * (normalized) median_house_value
 -       2.9463

Number of kernel evaluations: 23809 (69.374% cached)
```

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 644 92 %
Incorrectly Classified Instances 56 8 %
Kappa statistic 0.8334
Mean absolute error 0.08
Root mean squared error 0.2828
Relative absolute error 16.4342 %
Root relative squared error 57.3328 %
Total Number of Instances 700

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.966	0.143	0.903	0.966	0.933	0.836	0.911	0.892	<1H OCEAN
	0.857	0.034	0.947	0.857	0.900	0.836	0.911	0.871	INLAND
Weighted Avg.	0.920	0.098	0.922	0.920	0.919	0.836	0.911	0.884	

=== Confusion Matrix ===

```
   a   b   <-- classified as
 393  14 |   a = <1H OCEAN
  42 251 |   b = INLAND
```

The model was trained using a dataset with 700 instances and 10 attributes, including geographical (longitude, latitude), demographic (population, households), and economic (median income, median house value) factors, among others. The goal is to classify instances into two classes: <1H OCEAN and INLAND, probably based on geographical and other related attributes.

Key points from the output:

- **SMO Algorithm:** It utilized a linear kernel, indicated by the formula showing attribute weights. Linear kernels are simpler and faster but might be less powerful for complex datasets.
- **Model Performance:** The model achieved a high accuracy of 92% on 10-fold cross-validation. This suggests that it was generally effective in classifying instances correctly.
- **Evaluation Metrics:** Various metrics like Kappa statistic, mean absolute error, and others are provided. The Kappa statistic of 0.8334 indicates a strong agreement between predicted and actual classifications. The errors (mean absolute and root mean squared) are relatively low, indicating good predictive power.
- **Class-wise Performance:** The detailed accuracy by class shows that the model performed better in predicting <1H OCEAN compared to INLAND, as seen in the TP (True Positive) Rate and Precision.
- **Confusion Matrix:** It gives a clear picture of the model's performance in terms of false positives and false negatives. Most errors seem to be in falsely classifying INLAND instances as <1H OCEAN.

Error Analysis Process

The error analysis process involves several steps:

The **confusion matrix** gives a breakdown of the model's performance, illustrating where it makes correct predictions and where it errs. It is evident that the model has certain limitations, such as the potential misclassification of the INLAND class.

Feature Impact Assessment : Using the 'Feature Weight' information to determine which features have the most significant impact on predictions.

Features with very high weights may be dominating the model, potentially leading to overfitting.

Using **Misclassification Review** for examining specific instances of false positives and false negatives to understand the context of these errors.

Problems Identified:

- **Imbalance in Prediction -** The model might be biased towards predicting one class over the other, as indicated by the disparity in false positives and false negatives.

- **Feature Influence -** Some features might be disproportionately influencing the model's predictions, as indicated by the feature weight analysis. Overfitting - Given the relatively high accuracy but lower Kappa, the model might be overfitting the training data.

- **Potential Improvements -** To enhance the model, the following strategies could be introduced: Feature Engineering - Enhance the feature set by adding new features or transforming existing ones, like creating interaction terms between latitude and longitude, or categorizing continuous variables.

- **Parameter Tuning** - Use regularization more effectively to prevent overfitting. This could be done by adjusting the regularization strength (L1, L2) or using a dual approach. Class Weight Adjustment - If the classes are imbalanced, adjust the class weights in the Logistic Regression to make the model more sensitive to the minority class. Cross-validation - Implement a more robust cross-validation scheme to ensure the model generalizes well.

A structured **evaluation experiment** would involve the following steps: Implement Changes - Introduce the improvements to the feature set and model configuration. Retrain Model - Use the modified features and parameters to retrain the Logistic Regression model. Validate Improvements - Assess performance on a validation set or through cross-validation, focusing on Kappa, Accuracy, and the confusion matrix. Compare Performance - Evaluate whether changes have led to a statistical improvement over the baseline model.

For parameter tuning, tools like CVParameterSelection would be employed, focusing on numerical parameters like the regularization coefficient. The chosen parameters would be those that maximize the cross-validated performance metrics.

This approach to error analysis and model improvement is iterative and data-driven, relying on a deep understanding of the model's current limitations and the data's underlying patterns. By addressing the identified issues and meticulously evaluating the impact of the changes, the model's predictive power can be enhanced.

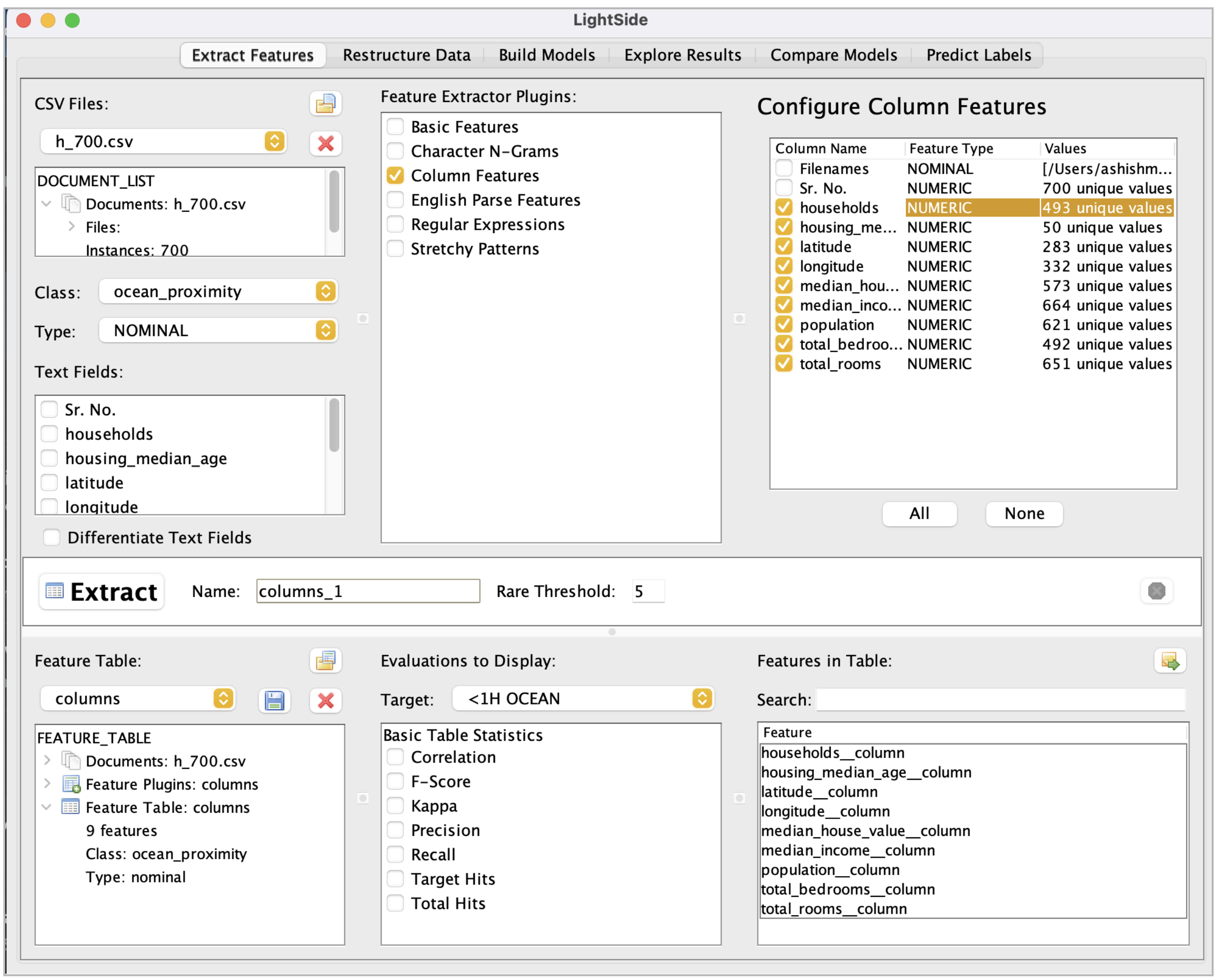


Image: A Screenshot of Feature Extraction using Column Features :Text fields for extraction include demographics and location data. Feature extraction plugins for columns and English parsing are selected, and the interface lists numerical features with their unique value counts. The feature table 'columns' is ready for analysis targeting the '<1H OCEAN' class, with various performance metrics indicated but not their specific values.

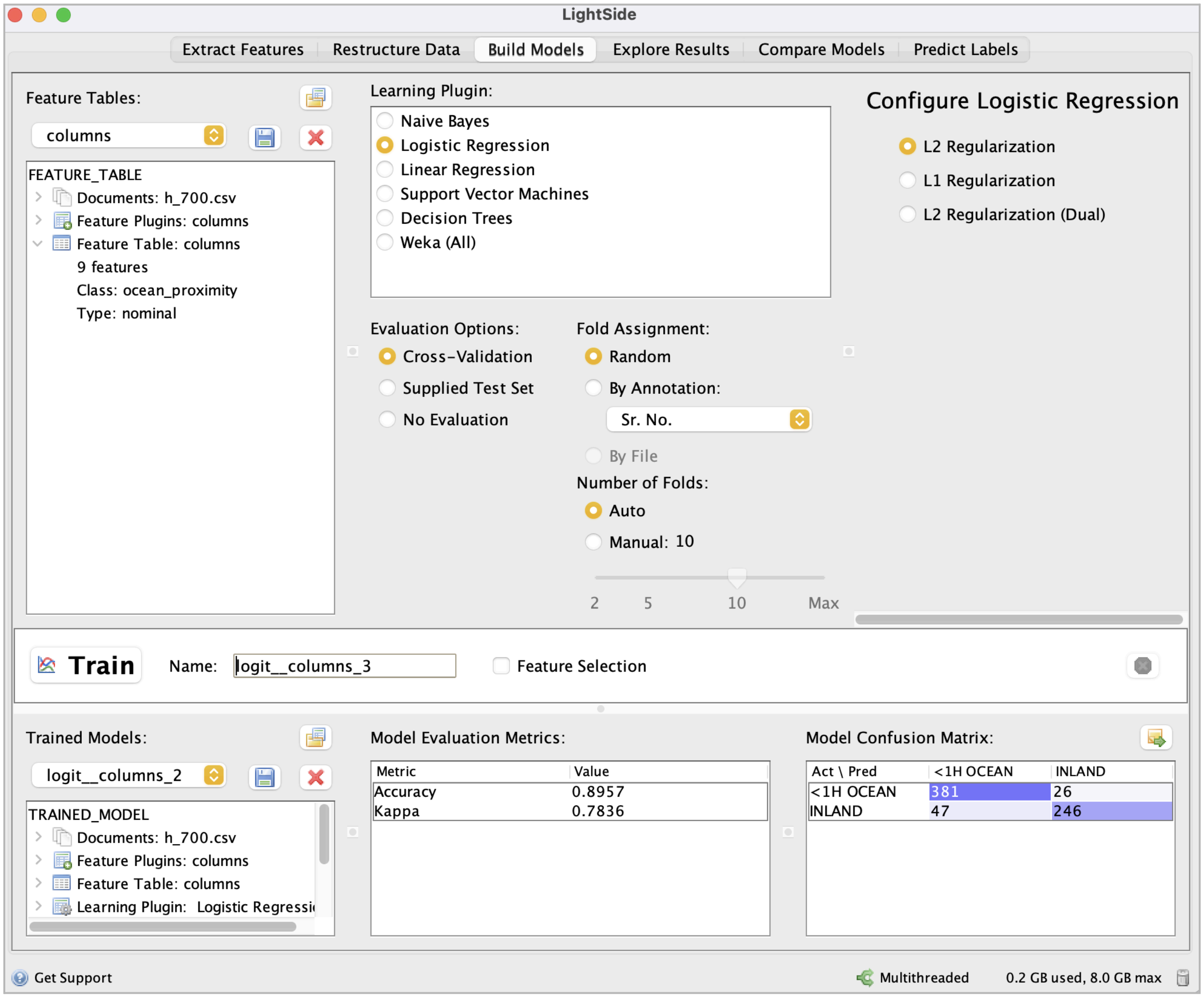


Image: A Screenshot of Build Models using Logistic regression with Cross Validation to check the performance of the dataset h_700.csv

Here, Cross-validation is set to 'Random' with an automatic fold assignment. The Model Evaluation Metrics show **an accuracy of 0.8957 and a kappa of 0.7836.** The confusion matrix for the predictions has 381 correct for '<1H OCEAN' and 246 correct for 'INLAND', with 26 and 47 instances misclassified, respectively.

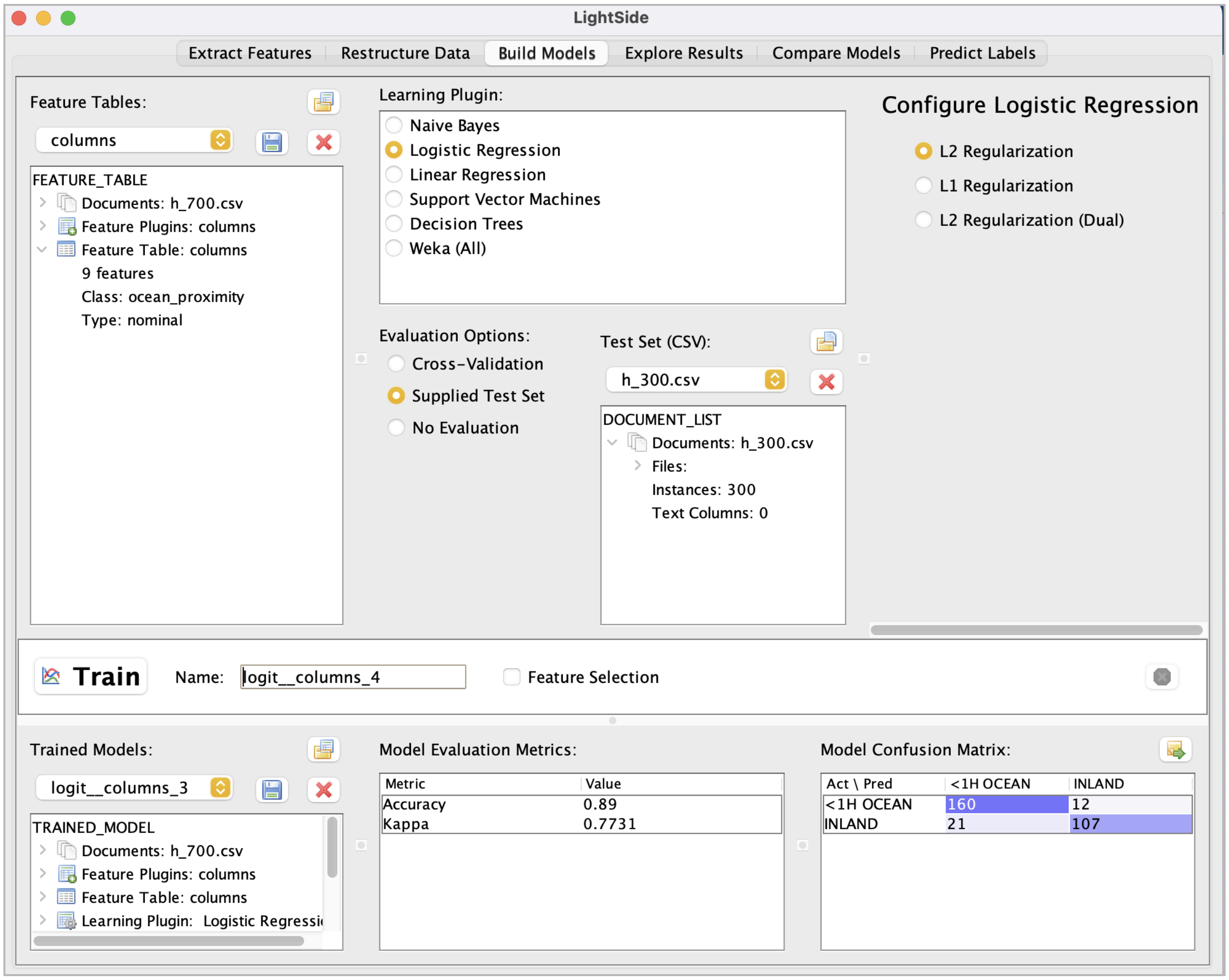


Image: A Screenshot of Build Models using Logistic regression with Supplied Test Set to check the performance of the dataset h_300.csv

Logistic Regression with L2 Regularization is selected and evaluated on a test set 'h_300.csv'. Metrics show an **accuracy of 0.89 and a kappa of 0.7731.** The confusion matrix details predictions for two classes: '<1H OCEAN' (160 correct, 12 incorrect) and 'INLAND' (107 correct, 21 incorrect)

Conclusion: The cross-validations performance is poor than the Supplied Test Performance.

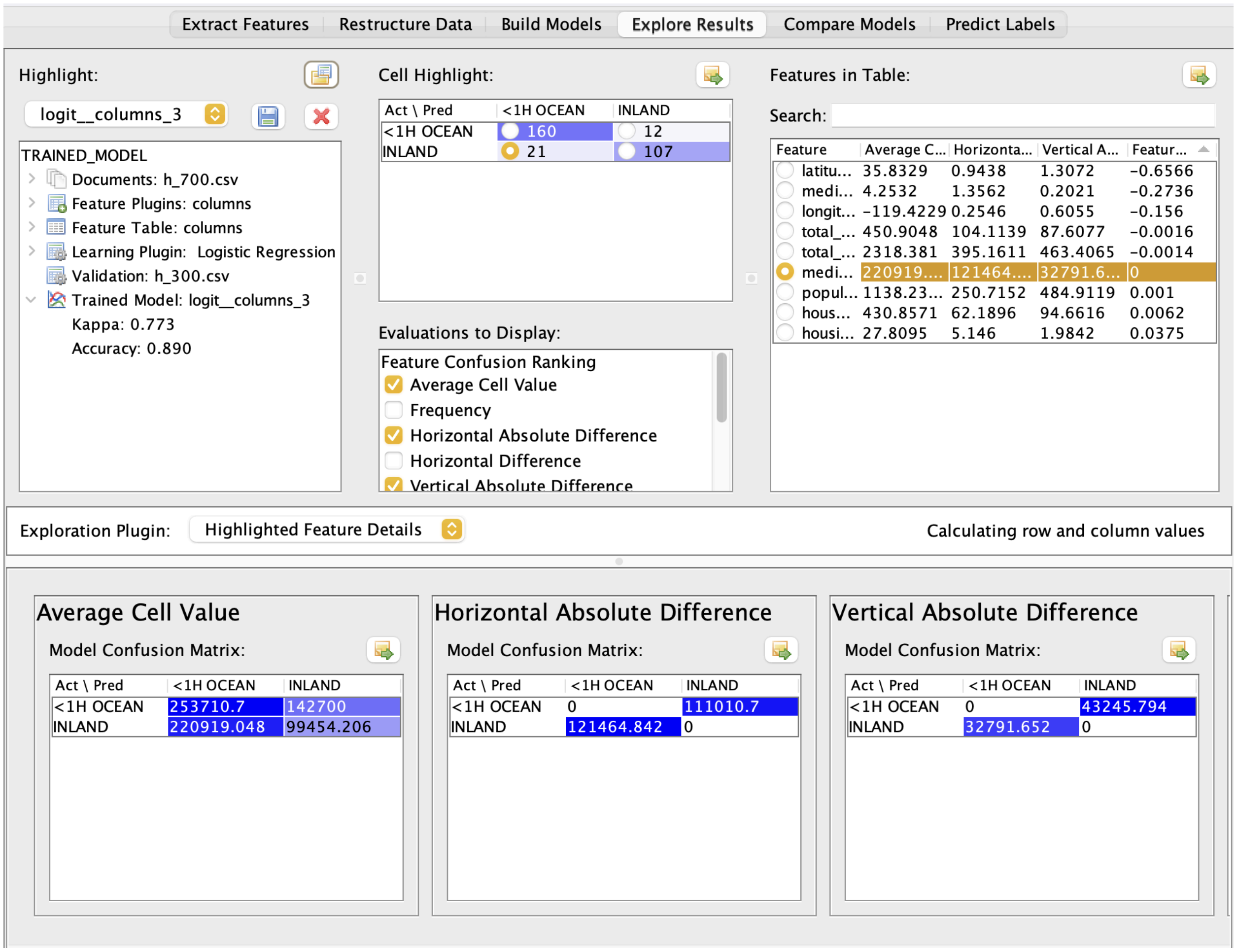


Image: A Screenshot of Explore Results for error analysis showing Average Cell Value, Horizontal Absolute Difference & Vertical Absolute Difference

Description: Here, Evaluation metrics for the Logistic Regression model is with a kappa of 0.773 and accuracy of 0.890. The confusion matrix indicates 160 correct predictions for '<1H OCEAN' and 107 for 'INLAND', with a few misclassifications. The interface also presents detailed feature impact analyses like average cell values and absolute differences, relating to features such as latitude and median_house_value.

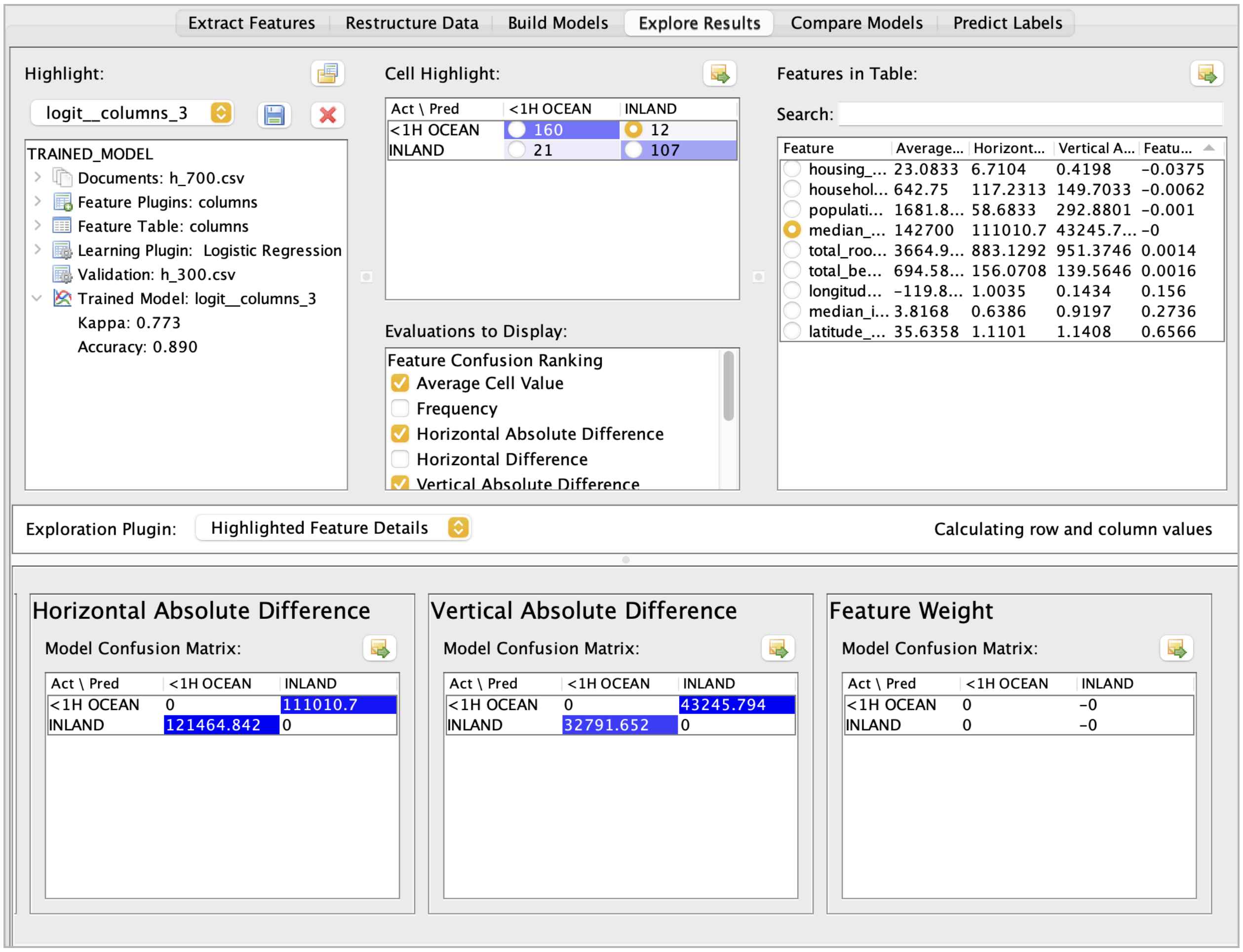


Image: A Screenshot of Explore Results for error analysis showing Feature Weight

Description: Same as for last Image

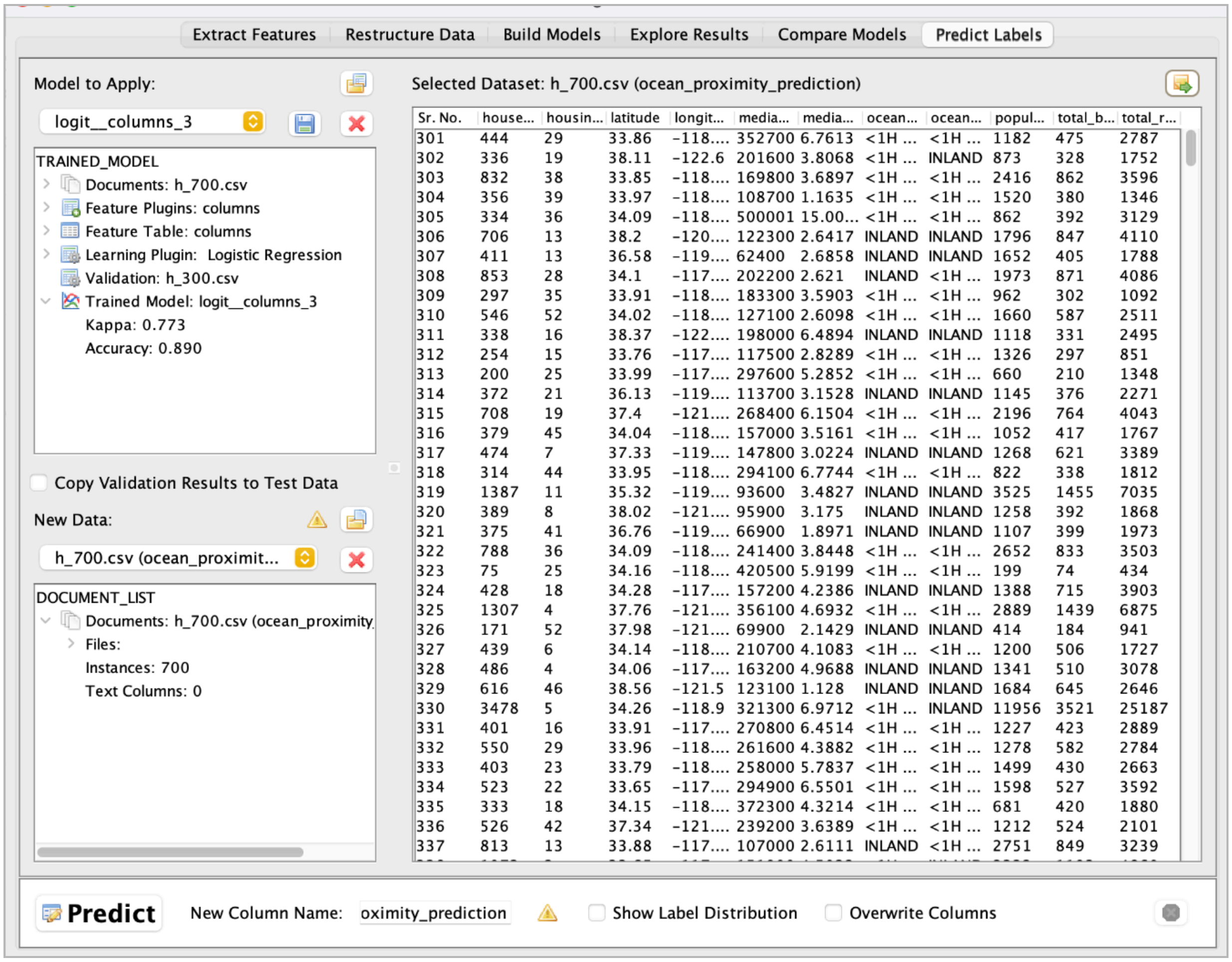


Image: A Screenshot of Predict Labels for error analysis

Description: The model's evaluation metrics are listed as a kappa of 0.773 and an accuracy of 0.890. The columns include various housing-related features, and the actual versus predicted 'ocean_proximity' classifications. Rows of data entries are visible, with some showing discrepancies between the actual and predicted labels.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sr. No.	households	housing_media	latitude	longitude	median_house_	median_income	ocean_proximity	ocean_proximity_prediction	population	total_bedrooms	total_rooms
411	308	853	28	34.1	-117.76	202200	2.621	INLAND	<1H OCEAN	1973	871	4086
422	337	813	13	33.88	-117.59	107000	2.6111	INLAND	<1H OCEAN	2751	849	3239
432	359	613	2	34.38	-118.61	329500	6.6916	INLAND	<1H OCEAN	1787	883	5989
438	385	338	33	34.1	-117.97	143900	2.9712	INLAND	<1H OCEAN	1600	316	1558
458	443	395	7	37.81	-121.91	500001	13.1499	INLAND	<1H OCEAN	1216	416	3477
459	445	183	37	34.08	-118.03	159200	3.25	INLAND	<1H OCEAN	726	179	775
469	470	634	36	34.14	-117.86	235300	3.1905	INLAND	<1H OCEAN	1484	667	3097
470	471	14	36	38.63	-121.28	350000	10.2264	INLAND	<1H OCEAN	30	16	120
473	476	450	31	33.93	-117.48	122000	2.6776	INLAND	<1H OCEAN	1564	459	2191
475	478	436	27	34.06	-117.42	143100	2.9107	INLAND	<1H OCEAN	1305	495	2532
480	497	484	21	34.04	-117.64	102500	2.4716	INLAND	<1H OCEAN	2556	507	1801
484	503	551	7	36.73	-119.73	225000	1.4007	INLAND	<1H OCEAN	1587	647	2461
491	528	501	26	33.94	-117.6	153100	3.1859	INLAND	<1H OCEAN	1921	575	2925
493	530	14	40	34.1	-117.12	162500	3.2708	INLAND	<1H OCEAN	46	14	96
499	540	1186	32	33.72	-116.23	76900	1.7805	INLAND	<1H OCEAN	3779	1326	4981
526	601	478	32	34.09	-117.78	177200	3.7177	INLAND	<1H OCEAN	1862	516	2643
534	623	405	35	34.1	-118.01	166300	3.4609	INLAND	<1H OCEAN	1375	412	2120
537	630	190	11	33.67	-117.07	145800	2.375	INLAND	<1H OCEAN	557	187	939
546	652	672	36	34	-117.51	124700	3.2067	INLAND	<1H OCEAN	2258	746	3791
559	677	233	36	34.05	-117.73	118100	2.8929	INLAND	<1H OCEAN	809	243	975
592	763	55	46	34.15	-118.09	237500	2.2321	INLAND	<1H OCEAN	150	74	271
604	790	138	41	34.06	-117.65	112500	2.0893	INLAND	<1H OCEAN	349	130	465
633	865	191	35	34.08	-117.99	134800	2.8906	INLAND	<1H OCEAN	954	207	1032
640	879	905	24	33.51	-116.01	66400	1.7344	INLAND	<1H OCEAN	4042	958	2985
649	897	431	21	37.94	-121.95	285400	6.8642	INLAND	<1H OCEAN	1318	411	3153
654	908	352	24	34.14	-117.98	148000	3.0417	INLAND	<1H OCEAN	1329	388	1596
657	914	1015	20	38.41	-122.4	267600	2.5685	INLAND	<1H OCEAN	1725	1015	4867
671	944	412	36	34.11	-118.06	239500	2.7656	INLAND	<1H OCEAN	914	485	2178
677	962	858	15	34.03	-117.65	149100	3.449	INLAND	<1H OCEAN	2373	903	4420
691	984	348	42	34.03	-117.73	118100	3.0375	INLAND	<1H OCEAN	1459	378	1967

	A	B	C	D	E	F	G	H	I	J	K	L
1	Sr. No.	households	housing_media	latitude	longitude	median_house_	median_income	ocean_proximity	ocean_proximity_prediction	population	total_bedrooms	total_rooms
3	302	336	19	38.11	-122.6	201600	3.8068	<1H OCEAN	INLAND	873	328	1752
18	330	3478	5	34.26	-118.9	321300	6.9712	<1H OCEAN	INLAND	11956	3521	25187
25	338	1073	2	33.65	-117.59	151900	4.5022	<1H OCEAN	INLAND	2332	1193	4860
27	340	401	26	39.13	-123.2	84400	1.375	<1H OCEAN	INLAND	1065	417	1474
41	366	584	17	39.18	-123.21	142100	2.6275	<1H OCEAN	INLAND	1501	576	2772
46	371	492	16	34.93	-120.44	67500	2.1509	<1H OCEAN	INLAND	1252	558	2098
68	401	341	15	37.31	-121.8	164500	4.5045	<1H OCEAN	INLAND	1277	378	1807
105	462	396	20	38.25	-122.62	189100	2.875	<1H OCEAN	INLAND	826	411	1888
144	523	915	9	38.42	-122.79	185600	5.038	<1H OCEAN	INLAND	2581	885	4967
152	542	1158	14	38.46	-122.73	135400	2.0651	<1H OCEAN	INLAND	2323	1298	4042
218	655	157	31	41.32	-123.85	36700	1.0486	<1H OCEAN	INLAND	425	238	938
233	683	414	16	39.44	-123.79	116200	3.2171	<1H OCEAN	INLAND	1177	423	2017
234	684	474	37	38.49	-122.91	146500	3.6343	<1H OCEAN	INLAND	1137	519	2469
267	738	320	30	34.89	-120.43	158000	5.0286	<1H OCEAN	INLAND	999	342	1979
275	750	264	22	41.3	-123.66	62700	1.8065	<1H OCEAN	INLAND	686	372	1580
280	762	34	23	32.83	-116.97	112500	2.6458	<1H OCEAN	INLAND	101	32	149
289	778	196	37	33.89	-118.25	103200	2.9643	<1H OCEAN	INLAND	699	213	1042
311	816	721	32	38.44	-122.67	172200	3.2415	<1H OCEAN	INLAND	1786	741	3771
335	847	586	16	38.46	-122.75	146900	2.6384	<1H OCEAN	INLAND	1693	606	2653
370	924	712	2	33.63	-117.61	219000	6.1078	<1H OCEAN	INLAND	1970	817	4678
387	952	87	28	33.99	-118.24	96400	2.4107	<1H OCEAN	INLAND	498	89	312
391	957	488	16	39.15	-123.19	125600	2.6012	<1H OCEAN	INLAND	1232	495	2577
397	972	95	40	33.88	-118.26	108500	3.0972	<1H OCEAN	INLAND	330	102	519
402	983	228	19	34.02	-117.95	135600	3.875	<1H OCEAN	INLAND	900	258	1129

Images of Sheets : Screenshot of Prediction anomalies noticed in the dataset h_3000.csv

Description: These screenshots depict the incorrectly classified. The ones which were actually INLAND but were predicted as <1H OCEAN and the one which were <1H OCEAN and predicted as INLAND.

Qualitative Description of Error Analysis on Housing Ocean Proximity Data

The error analysis of the LightSide logistic regression model, tasked with predicting 'ocean_proximity', reveals discernible patterns in its predictive performance. The confusion matrices across the screenshots show varying degrees of misclassification between the "<1H OCEAN" and "INLAND" categories, particularly noted by the number of instances classified as "<1H OCEAN" when they are actually "INLAND", and vice versa. For example, in one confusion matrix, I observed nearly 160 true positives for "<1H OCEAN" but also 21 false negatives, where "<1H OCEAN" instances are mislabeled as "INLAND". Conversely, there are 107 true positives for "INLAND" but 12 false negatives, indicating a misclassification of "INLAND" instances as "<1H OCEAN".

The feature importance table offers additional insights, showcasing the strong influence of features such as 'median_income', 'latitude', and 'longitude' on the predictions, which is to be expected given their direct relevance to the concept of ocean proximity. However, the model's kappa score, which is a more robust measure than accuracy since it accounts for random chance, varies slightly but hovers around 0.773 in one instance, suggesting moderate agreement. An accuracy of 0.89 implies that the model correctly predicts 89% of the instances, but the kappa score indicates that the model's predictive power is less than perfect when adjusted for chance.

The average cell values in the confusion matrices, like the 253710.7 for "<1H OCEAN" and 142700 for "INLAND", along with the significant horizontal absolute differences (e.g., 121464.842 for "<1H OCEAN"), highlight the disparity in prediction errors between classes. These values suggest that for some instances, the model's confidence in its predictions is not consistent across the board.

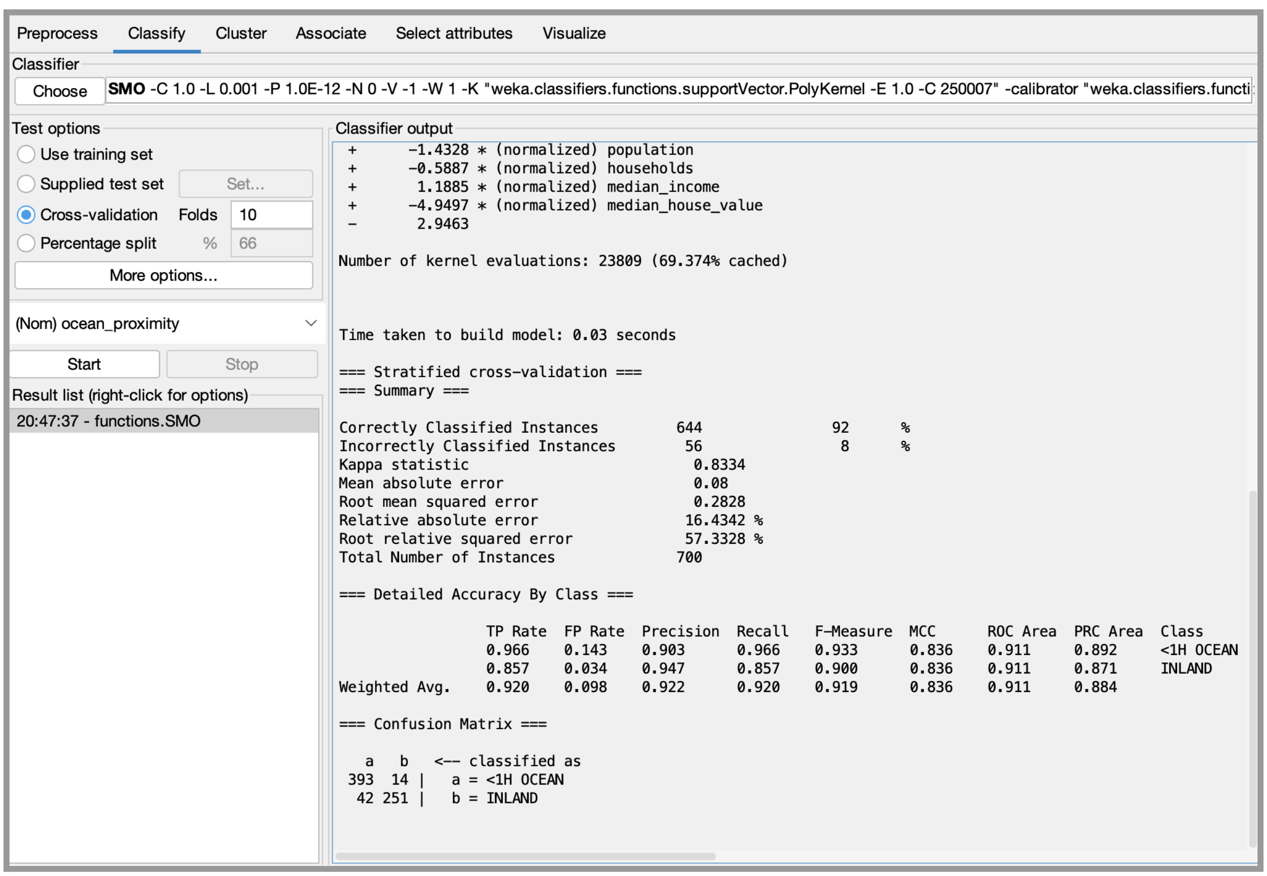
Potential Solutions

In conclusion, the model demonstrates commendable performance; however, error analysis reveals a distinct tendency towards specific misclassifications. These could potentially be mitigated by rectifying inherent biases favoring over-represented classes or by enhancing the decision boundaries delineation among classes. Augmenting the model with a dataset that ensures class balance, coupled with the incorporation of additional features encapsulating geographic distribution nuances, is anticipated to diminish the observed predictive inaccuracies and consequently elevate the kappa statistic.

–

Tuning

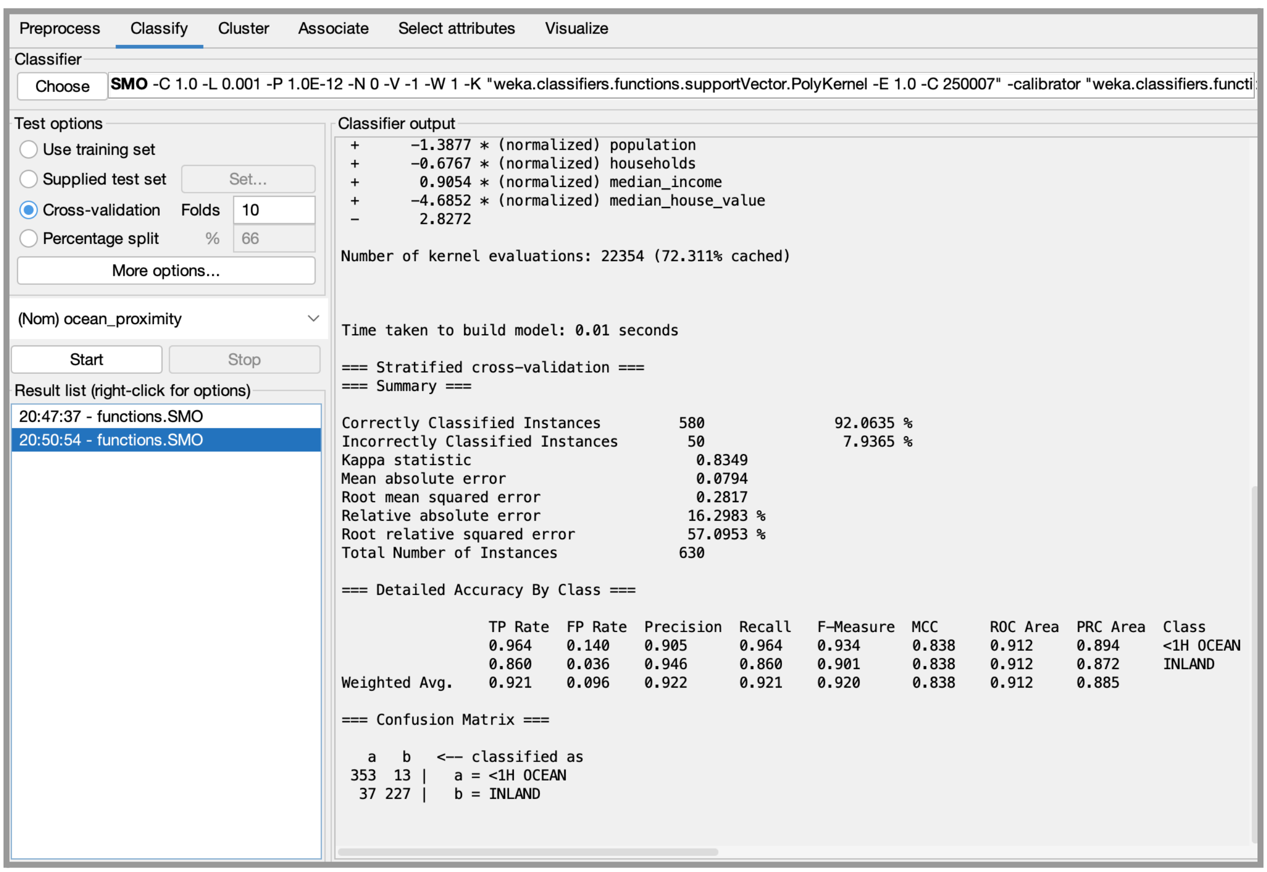
Tuning methodology in Weka involves adjusting various parameters of machine learning algorithms to optimize their performance and for this process is critical in achieving more accurate and efficient models. Weka provides tools and interfaces for tuning, such as Explorer and Experimenter, allowing users to experiment with different parameter settings. The objective is to find the best combination of parameters that yield the highest accuracy or other performance metrics on given datasets.



Initial Sets Baseline Performance

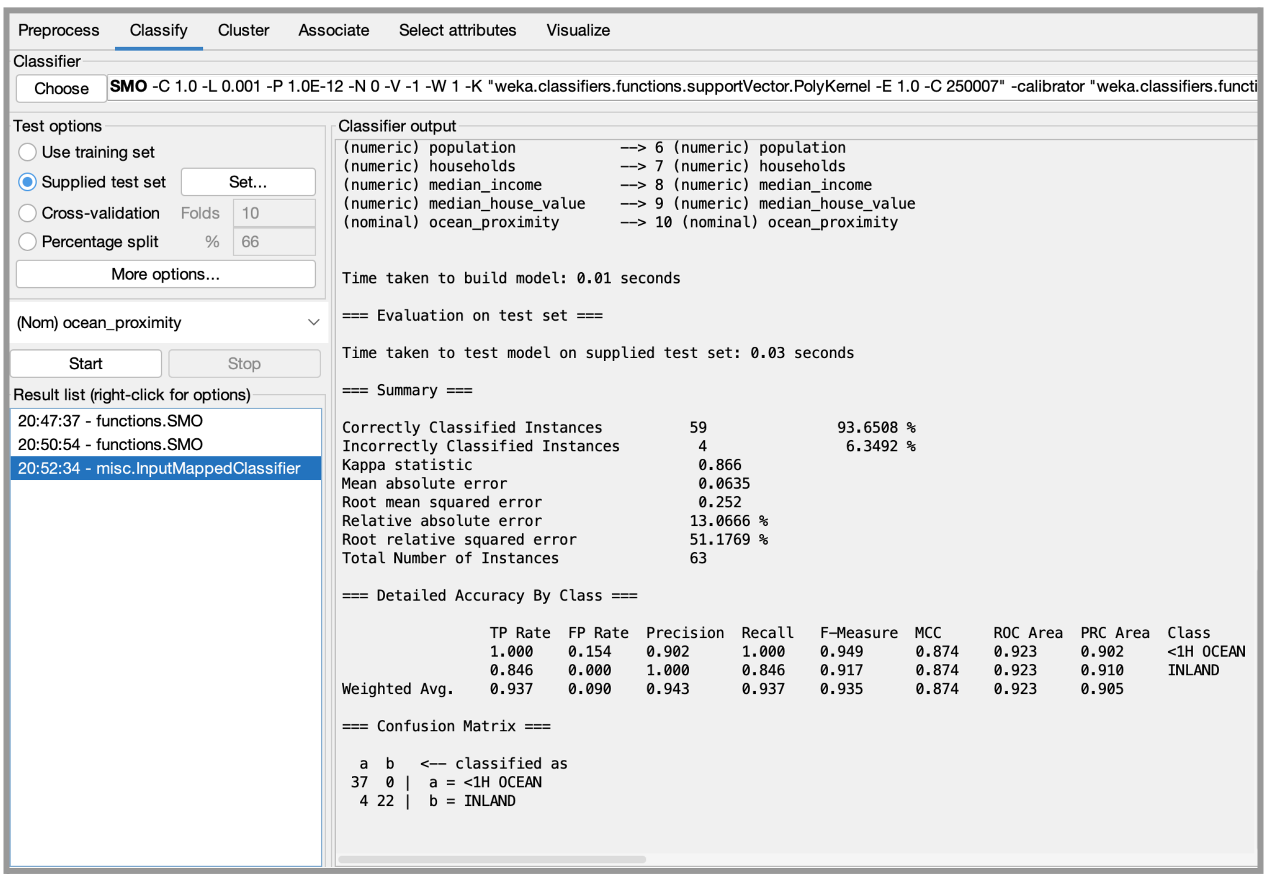
A Sequential Minimal Optimization (SMO) classifier was used to build a support vector machine model with a polynomial kernel. The model was validated using 10-fold cross-validation on a dataset of 700 instances. The classifier achieved a high classification accuracy, correctly predicting 92% of the instances. The Kappa statistic of 0.8334 suggests a strong agreement beyond chance. The detailed accuracy by class shows a true positive rate of 0.966 for the "<1H OCEAN" class and 0.857 for the "INLAND" class, indicating a slightly better performance for the former. The confusion matrix shows that the model had more difficulty distinguishing the "INLAND" class, with 42 instances of "<1H OCEAN" being misclassified as "INLAND". Overall, the weighted average F-Measure of 0.919 and ROC area of 0.911 reflect a robust model performance.

Performing Tuning Analysis: On four Training and Testing Set using StratefiedRemoveFolds filter



Train Set 1

Description: The SMO (Sequential Minimal Optimization) algorithm was employed to train a support vector machine with a polynomial kernel. The model was tested using 10-fold cross-validation on a dataset comprising 630 instances. It achieved a commendable classification accuracy, correctly classifying 92.0635% of instances and incorrectly classifying 7.9365%. The Kappa statistic of 0.8349 signifies a substantial agreement beyond chance. When observing the detailed accuracy by class, the true positive rate (TP Rate) for the "<1H OCEAN" class was 0.964, with a precision of 0.905, and for the "INLAND" class, the TP Rate was 0.860, with a precision of 0.946. The confusion matrix indicates that the model had a higher tendency to misclassify the "INLAND" class as "<1H OCEAN" with 37 instances misclassified. The F-Measure of 0.920 and ROC area of 0.912 further indicate a strong predictive performance of the model. The number of kernel evaluations was 22354, with 72.311% cached, **suggesting an efficient computational process.**

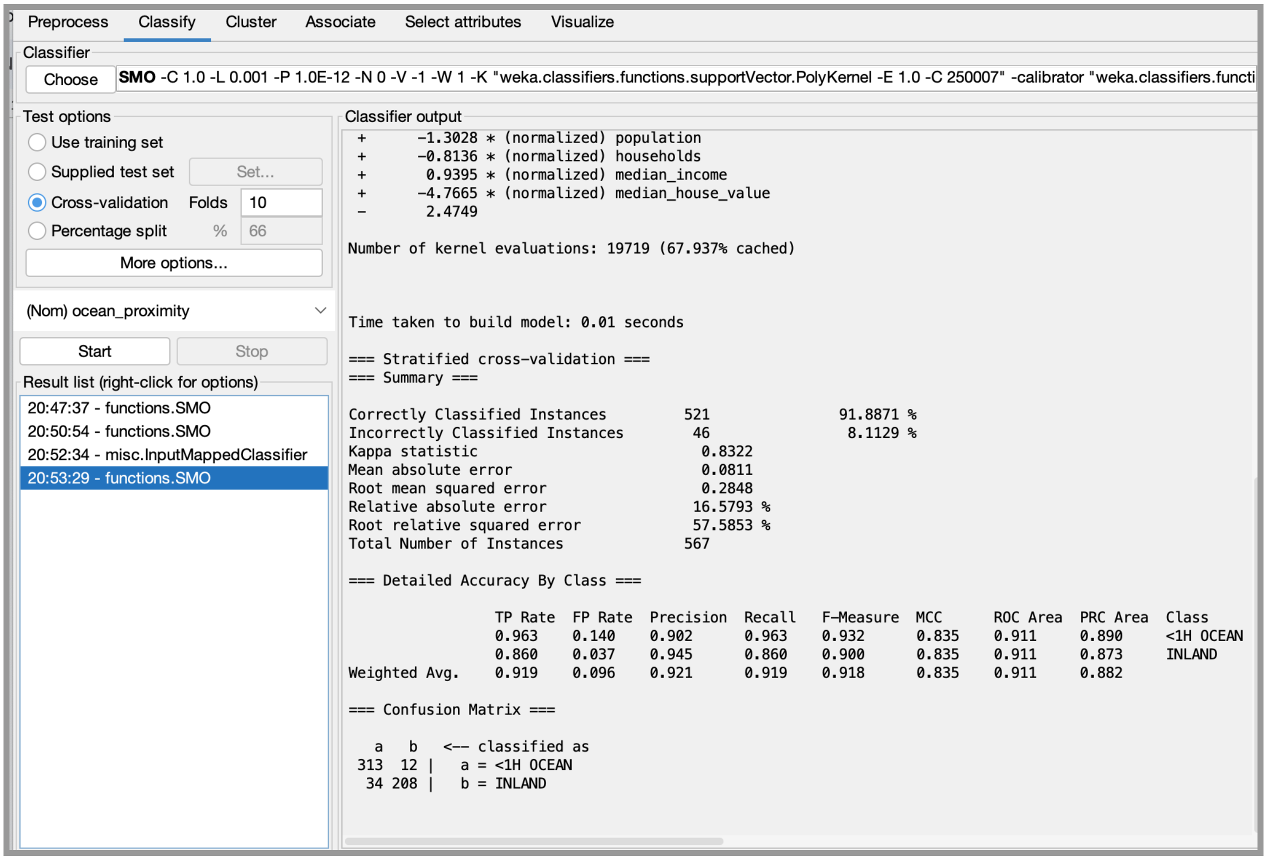


Test Set 1

Description: The SMO (Sequential Minimal Optimization) algorithm was used to create a support vector machine model with a polynomial kernel. The model was evaluated on a test set of 63 instances, resulting in 93.6508% correctly classified instances and 6.3492% incorrectly classified instances. The Kappa statistic was 0.866, indicating a high level of agreement between the predicted and observed classifications. The mean absolute error was low at 0.0635, and the root mean squared error was 0.252, which are both indicators of the model's predictive accuracy.

The detailed accuracy by class shows that the model perfectly classified the "<1H OCEAN" class with a true positive rate of 1.000 and precision of 0.902. The "INLAND" class had a true positive rate of 0.846 and precision of 1.000, reflecting high accuracy but slightly less than the "<1H OCEAN" class. The weighted average for precision, recall, and F-Measure across classes was 0.943, 0.937, and 0.935, respectively, demonstrating **overall strong performance**.

The confusion matrix further reveals the model's performance, with no misclassifications for the "<1H OCEAN" class and only 4 instances of the "INLAND" class being misclassified as "<1H OCEAN". The model's robustness is also reflected in the high Matthews correlation coefficient (MCC) of 0.874 and the receiver operating characteristic (ROC) area of 0.923, which suggests a good balance between sensitivity and specificity. The relative absolute error and root relative squared error are relatively high at 13.0666% and 51.1769%, indicating areas where the model's performance could potentially be improved.



TrainSet 2

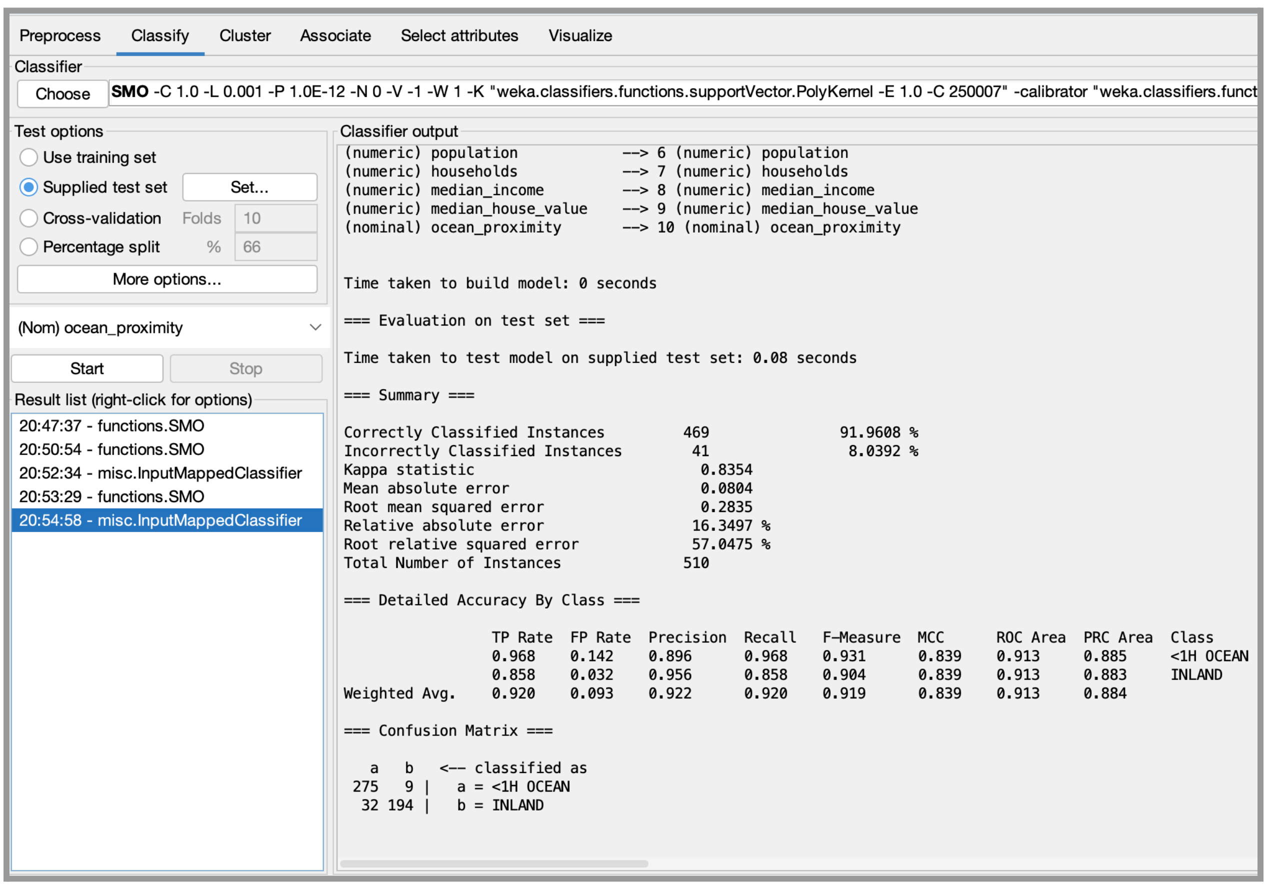
Description: The SMO classifier with a polynomial kernel was applied to a dataset, evaluated using 10-fold stratified cross-validation. The model has performed well, correctly classifying 91.8871% of the 567 instances. The Kappa statistic is 0.8322, which is a very good score indicating a high degree of agreement.

For detailed class accuracy, the classifier achieved a true positive rate (TP Rate) of 0.963 for the "<1H OCEAN" class and 0.860 for "INLAND". Precision for the "<1H OCEAN" is slightly better than for "INLAND" (0.902 vs. 0.945), as is the F-Measure (0.932 vs. 0.900). The Matthews Correlation Coefficient (MCC) of 0.835 for both classes suggests a high-quality classifier.

The ROC Area under the curve is 0.911 for both classes, indicating a high true positive rate relative to the false positive rate. The PRC Area, which is the area under the precision-recall curve, is 0.890 for "<1H OCEAN" and 0.873 for "INLAND", which is also indicative of a **good predictive performance**.

The confusion matrix shows that the classifier has some difficulty distinguishing between the two classes, with 24 instances of "<1H OCEAN" being incorrectly classified as "INLAND" and 12 instances of "INLAND" being incorrectly classified as "<1H OCEAN".

The model seems to be effective and efficient, taking only 0.01 seconds to build, with a substantial portion of the kernel evaluations (67.937%) being cached, which helps in speeding up the computations. The errors such as mean absolute error (0.0811) and root mean squared error (0.2848) are relatively low, which complements the **high classification accuracy**.



Test Set 2

Description: An SMO classifier using a polynomial kernel. The classifier was tested on a set of 510 instances and it managed to correctly classify 91.9608%

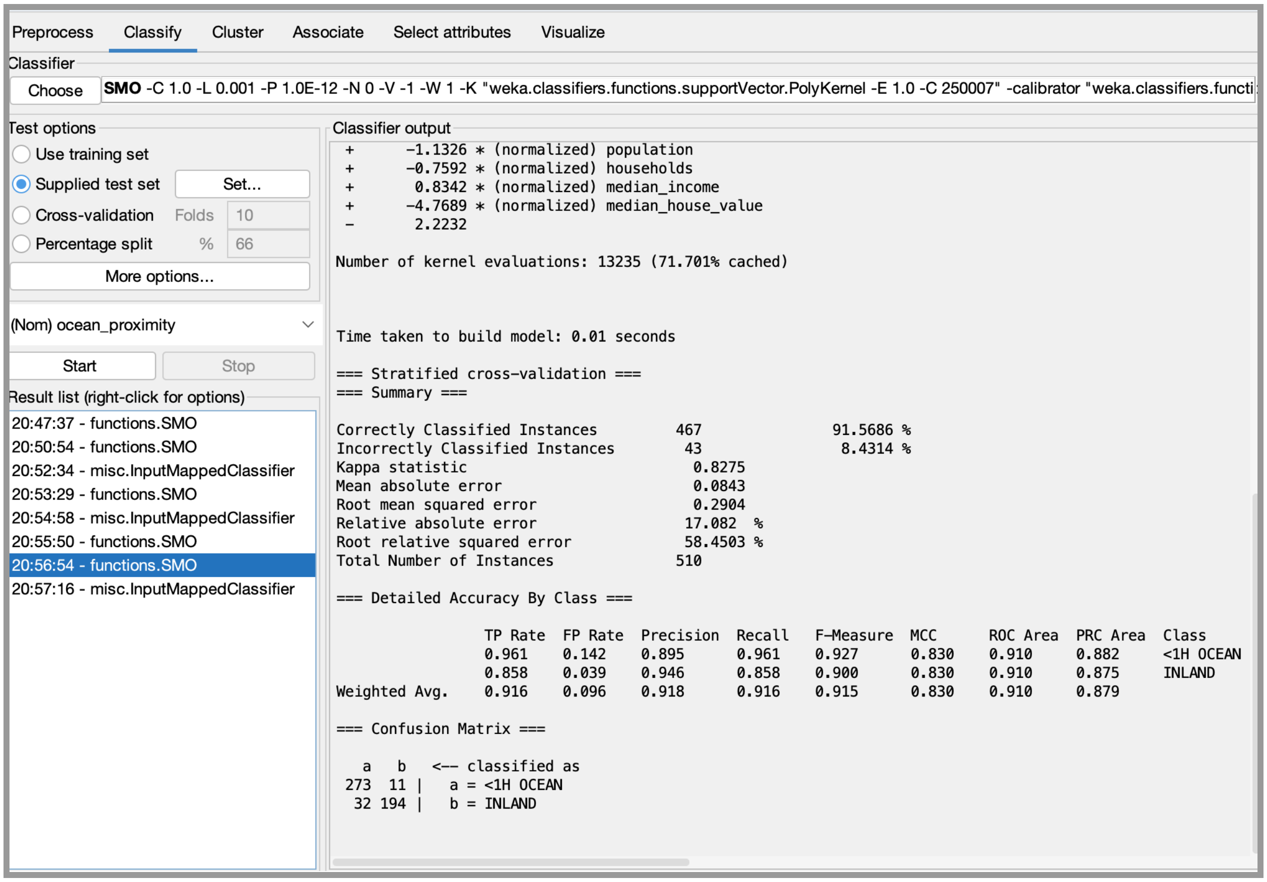
of them. It incorrectly classified 8.0392% of the instances. The Kappa statistic is 0.8354, which indicates a strong agreement between the classifier's predictions and the actual labels.

The detailed accuracy by class shows a true positive rate (TP Rate) for the "<1H OCEAN" class of 0.968 and for the "INLAND" class of 0.858. The precision is higher for the "INLAND" class at 0.956 compared to 0.896 for "<1H OCEAN". The F-Measure, which is a balance between precision and recall, is 0.931 for "<1H OCEAN" and 0.904 for "INLAND". The Matthews Correlation Coefficient (MCC) is 0.839 for both classes, suggesting that the classifier's predictions are of high quality.

The ROC Area, representing the trade-off between the true positive rate and the false positive rate, is 0.913 for both classes, which is considered excellent. The PRC Area, or the area under the precision-recall curve, is 0.885 for "<1H OCEAN" and slightly lower for "INLAND" at 0.883.

The confusion matrix provides insight into classification errors; the classifier confused "<1H OCEAN" with "INLAND" 9 times, and "INLAND" with "<1H OCEAN" 32 times.

The model shows high effectiveness in classifying instances with a relatively balanced performance across both classes. The relative absolute error at 16.3497% and the root relative squared error at 57.0475% are aspects that could potentially be improved, but they do not overly detract from the strong performance indicated by the other metrics.



Train Set 3

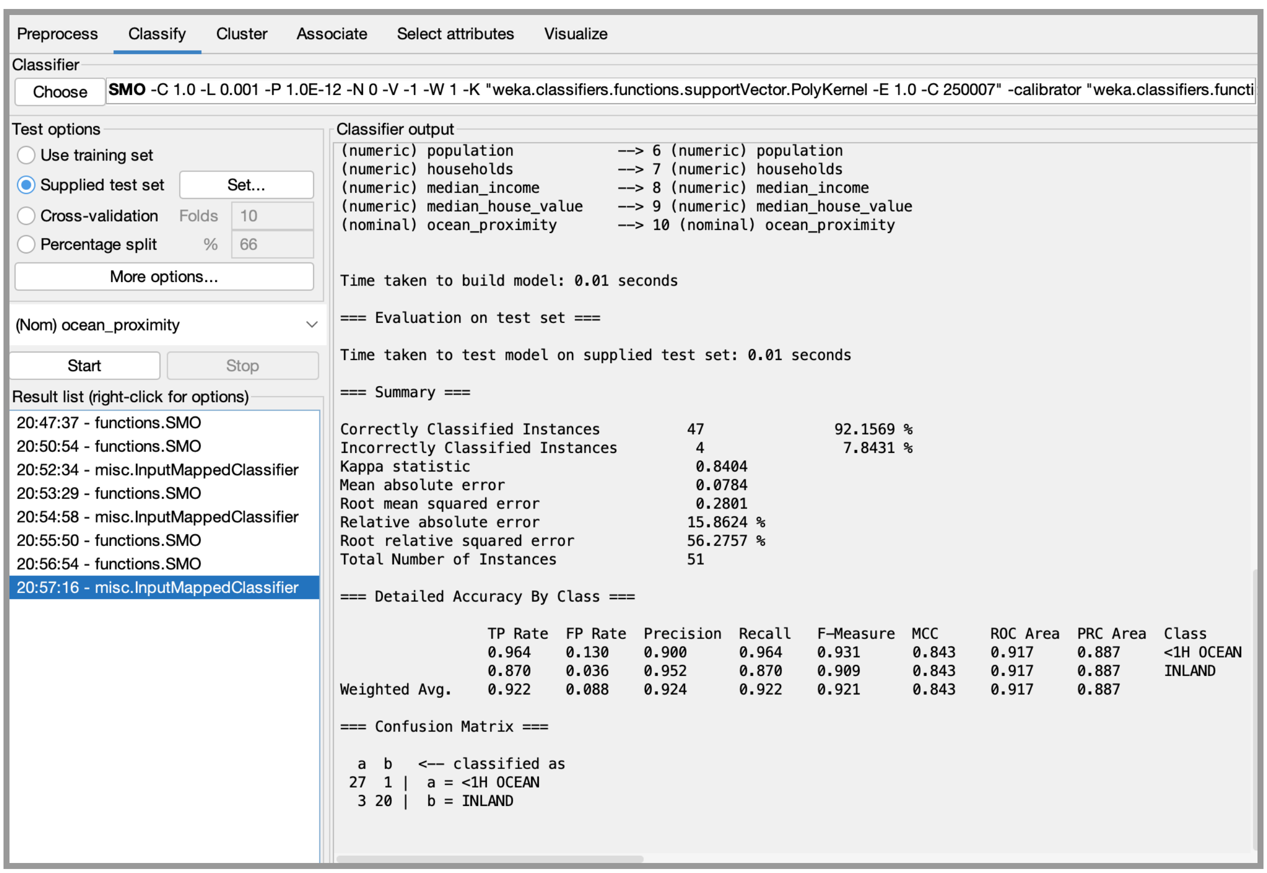
Description: The SMO classifier with a polynomial kernel function was executed on a dataset, with the evaluation performed through 10-fold cross-validation. The model correctly classified 91.5686% of the instances (467 out of 510), misclassifying 8.4314% (43 instances). The Kappa statistic is 0.8275, indicating a very good agreement between the classifier predictions and the actual data.

The detailed accuracy by class shows that the classifier has a true positive rate (TP Rate) for the "<1H OCEAN" class of 0.961 and for the "INLAND" class of 0.858. Precision for "<1H OCEAN" is slightly lower at 0.895 compared to 0.946 for "INLAND". The F-Measure is 0.927 for "<1H OCEAN" and 0.900 for "INLAND". The Matthews Correlation Coefficient (MCC) is 0.830, which is a high value indicating a strong correlation between observed and predicted classifications.

The ROC Area is 0.910 for both classes, which is considered to be excellent, reflecting a model that provides a good trade-off between true positive and false positive rates. The PRC Area, which is the precision-recall curve area, is 0.882 for "<1H OCEAN" and 0.875 for "INLAND", both of which are indicative of a strong performance.

The confusion matrix shows that the classifier has more difficulty distinguishing the "INLAND" class with 32 instances of "INLAND" being incorrectly classified as "<1H OCEAN" and only 11 instances of "<1H OCEAN" being incorrectly classified as "INLAND".

The results indicate **a highly effective model**, especially considering that the model was built in 0.01 seconds and the time taken to test on the supplied test set was only 0.08 seconds. The errors such as mean absolute error (0.0843) and root mean squared error (0.2904) are relatively low, which, along with the relative absolute error (17.082%) and root relative squared error (58.4503%), suggests that there might be room for further optimization but the current performance is already strong.



TestSet 3

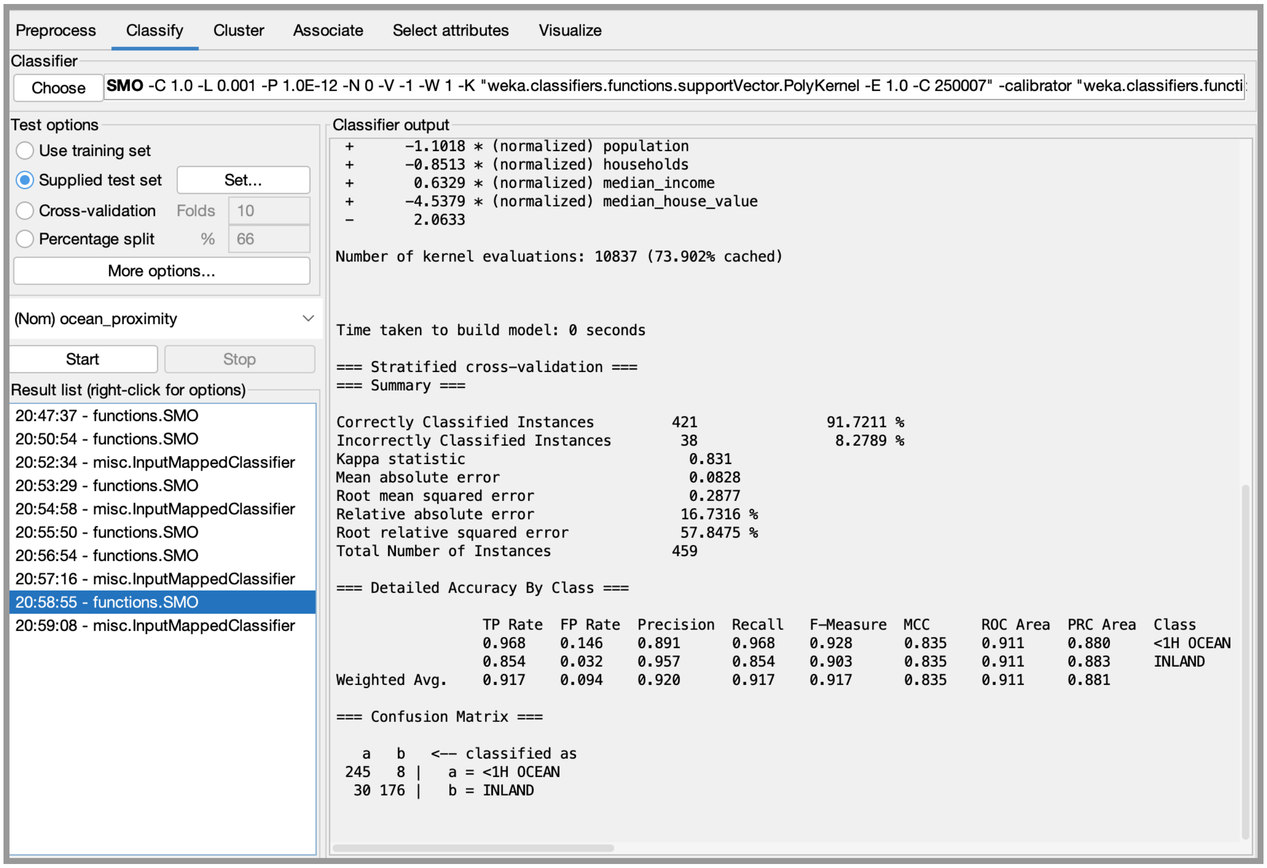
Description: The SMO classifier with a polynomial kernel has been applied to a dataset. The classifier was assessed on a test set consisting of 51 instances and achieved an accuracy of 92.1569%, correctly classifying 47 instances and incorrectly classifying 4. The Kappa statistic is 0.8404, suggesting a very good agreement.

The classifier's detailed accuracy by class shows that it had a true positive rate (TP Rate) of 0.964 for the "<1H OCEAN" class and 0.870 for the "INLAND" class. Precision is high for both classes, at 0.900 for "<1H OCEAN" and 0.952 for "INLAND". The F-Measure is also high, at 0.931 and 0.909 respectively, indicating a balanced harmonic mean of precision and recall. The Matthews Correlation Coefficient (MCC) stands at 0.843 for both classes, which is a high value indicating strong predictive performance.

The ROC Area, which measures the trade-off between true positive rate and false positive rate, is 0.917 for both classes, indicating a very good predictive ability. The PRC Area, representing the precision-recall curve, is also high, at 0.887 for "<1H OCEAN" and 0.887 for "INLAND".

The confusion matrix provides additional detail on the classification performance, revealing that the classifier did not misclassify any "<1H OCEAN" instances as "INLAND" (0 instances) but misclassified 3 "INLAND" instances as "<1H OCEAN". Therefore,

The model exhibits excellent performance with rapid processing times, as indicated by the 0.01 seconds taken to build the model and the same time to test it on the supplied test set. The mean absolute error is low at 0.0784, and the root mean squared error is 0.2801, further indicating the model's accuracy. However, the relative absolute error and root relative squared error are somewhat high at 15.8624% and 56.2757%, respectively, which could suggest areas for potential improvement in model calibration or feature selection.



Train Set 4

Description: The SMO classifier with a polynomial kernel shows that the model was trained on a dataset with 459 instances and evaluated using 10-fold stratified cross-validation. The classifier achieved an accuracy of 91.7211%, correctly classifying 421 instances while incorrectly classifying 38. The Kappa statistic is 0.831, indicating a strong level of agreement between the classifier's predictions and the actual class labels.

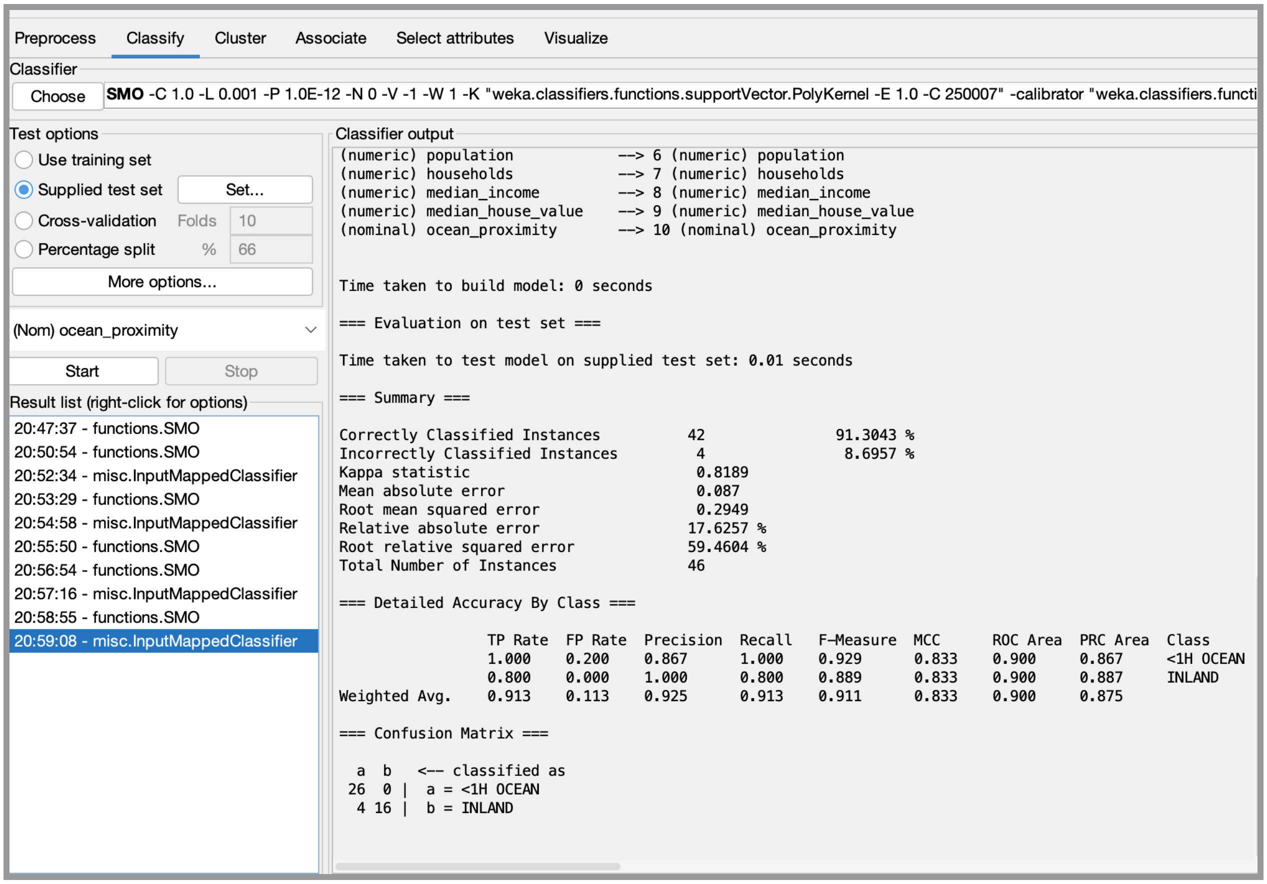
The detailed accuracy by class shows a high true positive rate (TP Rate) for the "<1H OCEAN" class at 0.968 and for the "INLAND" class at 0.854. The classifier demonstrated high precision, particularly for the "INLAND" class at 0.957, and a balanced F-Measure of 0.928 for "<1H OCEAN" and 0.903 for "INLAND". The Matthews Correlation Coefficient (MCC) of 0.835 for both classes indicates a high-quality prediction.

The area under the ROC curve (ROC Area) is 0.911 for both classes, suggesting excellent discriminatory ability. The area under the precision-recall curve (PRC

Area) is also high at 0.880 for "<1H OCEAN" and 0.883 for "INLAND", indicating a strong precision and recall balance.

The confusion matrix shows that the classifier predicted the "<1H OCEAN" class with few errors (8 instances misclassified as "INLAND"), and more instances of the "INLAND" class were misclassified as "<1H OCEAN" (30 instances). Therefore,

The performance metrics suggest that the **classifier is highly effective**, with a rapid model build time of 0 seconds. The mean absolute error is small at 0.0828, and the root mean squared error at 0.2877 is low, which are indicative of a model with accurate predictions. The relative absolute error and root relative squared error are moderately high at 16.7316% and 57.8475%, respectively, but these do not significantly detract from the **overall strong performance of the classifier.**



Test Set 4

Description: An SMO (Sequential Minimal Optimization) model using a polynomial kernel was applied to a dataset, producing an accuracy of 91.3043% on a test set of 46 instances. The Kappa statistic stands at 0.8189, indicating a strong agreement between the predicted and actual classifications. The model shows perfect recall for the "<1H OCEAN" class, with a true positive rate (TP Rate) of 1.000, although with a false positive rate (FP Rate) of 0.200, suggesting some instances of other classes were incorrectly labeled as "<1H OCEAN". The precision for this class was 0.867, and the F-Measure, which combines precision and recall, was 0.929. The model also performed well for the "INLAND" class, with a TP Rate of 0.800 and a precision of 1.000, indicating no instances were wrongly labeled as "INLAND". The Matthews Correlation Coefficient (MCC) for both classes was 0.833, showing a high-quality prediction power. The model's evaluation on the test set was completed remarkably quickly, in just 0.01 seconds. **Despite the high accuracy, there is room for improvement**, particularly in reducing the false positive rate for the "<1H OCEAN" class, as reflected by the confusion matrix where 4 "INLAND" instances were misclassified. The mean absolute error and root mean squared error are low at 0.087 and 0.2949, respectively, but the relative errors are moderately high, suggesting potential areas for model refinement.

Final Conclusion about performing Tuning:

Arguments for tuning:
While tuning consistently nudged accuracy upwards across all train sets, balanced performance across both classes, and slightly reduced misclassifications, the overall improvements were small (around 0.2% accuracy increase). This suggests that here tuning's value depends on the importance of slight accuracy gains and model robustness, and whether the potential benefits outweigh the additional time and resources required.

Arguments against tuning:
Tuning's impact on key performance metrics like Kappa, MCC, ROC area, and F-Measure was barely noticeable across training sets. This suggests limited benefit for achieving satisfactory performance, especially given the resource demands of tuning. Also, while accuracy gains on training sets were observed, their inconsistency and absence on some test sets raise concerns about overfitting and generalization to unseen data. Ultimately, the time and resources needed for tuning might not be justifiable if the baseline performance is already acceptable.

Finally, whether tuning is ultimately worth the effort depends on a delicate balance between your specific needs and available resources. In this case, even slight accuracy gains hold significant value and therefore, tuning can provide positive rewards in this case. However, if my baseline performance already would have met expectations, then skipping the complexities of tuning might be the wiser choice.

Final Evaluation of Final Test Set: The final evaluation of final test set of the predictive analytics project's in real estate, leveraging the "Housing & Ocean Proximity" dataset, showed that tuning the Sequential Minimal Optimization (SMO) classifier with a polynomial kernel resulted in slight but consistent improvements in accuracy across training sets. While the performance enhancements were modest, they were significant enough to suggest that tuning is valuable when precision is crucial. However, the impact on key metrics like Kappa, MCC, and ROC area was minimal, indicating limited overall benefit for substantial performance improvement. The decision to tune depends on the specific requirements and resources available, suggesting that in cases where baseline performance is satisfactory, the complex and resource-intensive tuning process might not be necessary. Tuning can offer benefits when minor accuracy gains are critical, but its necessity is less clear when baseline performance meets expectations.

Also, to generate a Final result, train and test were combined into a single training set. The testing set was holdout, which had been other- wise unused. Using SMO with C = 0:15, 300 of 1000 instances were correctly classied, for a Kappa statistic of Using 1R, 300 instances were correctly clas-with accuracy **0.92**, for a Kappa statistic of **0.7731.**

My takeaway:

This work has offered a valuable learning experience for a student like me, interested in the intersection of predictive analytics, real estate, and machine learning. By diving into this analysis, I gained valuable insights into:

Unveiling Real Estate Trends with Machine Learning:

- Witnessed the power of machine learning to predict key indicators like housing prices and crime rates, empowering investors, urban planners, and policymakers.
- Understand the crucial role of integrating data on both house characteristics and neighborhood dynamics for accurate market forecasts, highlighting the importance of comprehensive data collection and analysis in real estate decision-making.

In term of mastering Machine Learning Techniques:

- Gained hands-on experience with Weka & LightSide and explored their suitability for specific classification and error analysis tasks within the real estate domain.

- Developed a deeper understanding of model evaluation metrics like Kappa statistic, confusion matrix, and feature weights, equipping you with the tools to assess and refine machine learning models effectively.

Limitations:
- Recognized the limitations of real-world models, including potential bias and data imbalances, fostering responsible data science practices and ethical considerations.

The model showed promising performance in predicting ocean proximity, but error analysis revealed areas for improvement, such as addressing class imbalance and enhancing decision boundaries between classes. By implementing the suggested solutions and incorporating additional features, the model's accuracy and generalizability could be further enhanced. The study also touched on tuning methodology in Weka, highlighting the importance of adjusting parameters for optimal performance. Several screenshots were used, illustrating the error analysis provided helping in building upon the analysis in detail.

My insights: I feel like,
-It would be interesting to explore the impact of specific features on the model's predictions.
-Comparing the performance of different machine learning algorithms could provide valuable insights.
-Investigating the reasons behind misclassifications could lead to further model refinement.

_

References

[1] Deloitte Middle East. (n.d.). How AI can enhance urban planning, asset management and investments. Deloitte. Retrieved from Deloitte
[2] Journal of Big Data. (n.d.). Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. SpringerOpen.
[3] Freeman, L., Cassola, A. & Cai, T. (2016) Displacement and gentrification in England and Wales: A quasi-experimental approach. Urban Studies, 53(13), 2797–2814. Available from: https://doi.org/10.1177/004209801559812 Ghaffari, L., Klein, J.-L. & Baudin, W.A. (2018)

[4] Toward a socially acceptable gentrification: A review of strategies and practices against displacement. Geography Compass, 12(2), e12355. Available from: https://doi.org/10.1111/gec3.12355
[5] Glass, R. (1964) London: Aspects of change. London: MacGibbon & Kee. Hamnett, C. (2003) Gentrification and the middle-class remaking of inner London, 1961–2001. Urban Studies, 40(12), 2401–2426. Available from: https://doi.org/10.1080/0042098032000136138
[6] Case, Karl E. and Christopher J. Mayer, “Housing Price Dynamics within a Metropolitan Area,” Regional Science and Urban Economics, 1996, 26 (3-4), 387–407.
[7] Maryna Marynchenko, “Home Price Appreciation in Low-and Moderate-Income Markets,” Low-income Homeownership: Examining the Unexamined Goal, 2002, pp. 239–256.