New Development Property Prediction

Property Advertisement Description Classification

Domain

Supervisor: Min Li
(Head of Data Science - Data Quality and Information Extraction)
Major: Master of Data Science
Name : Lei Yang

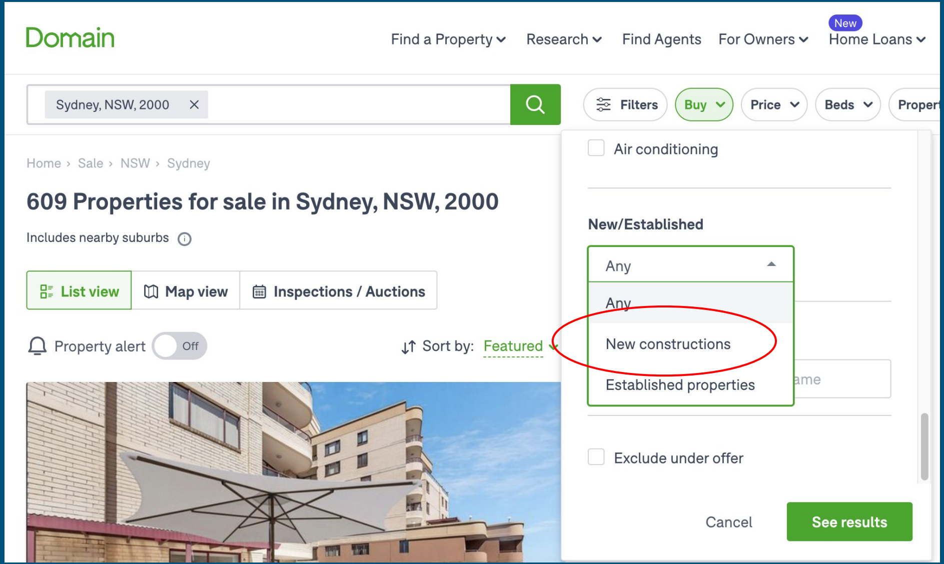
PROJECT OBJECTIVE

Project Title: Description Classification - New Development vs. Established

Project Objective/Goal: Assist the business in understanding the number of new development properties in the market

Problem Description:

- Improve filtering accuracy
- Improve user experience
- Enhance scalability of profit



Sydney, NSW, 2000

Home › Sale › NSW › Sydney

609 Properties for sale in Sydney, NSW, 2000

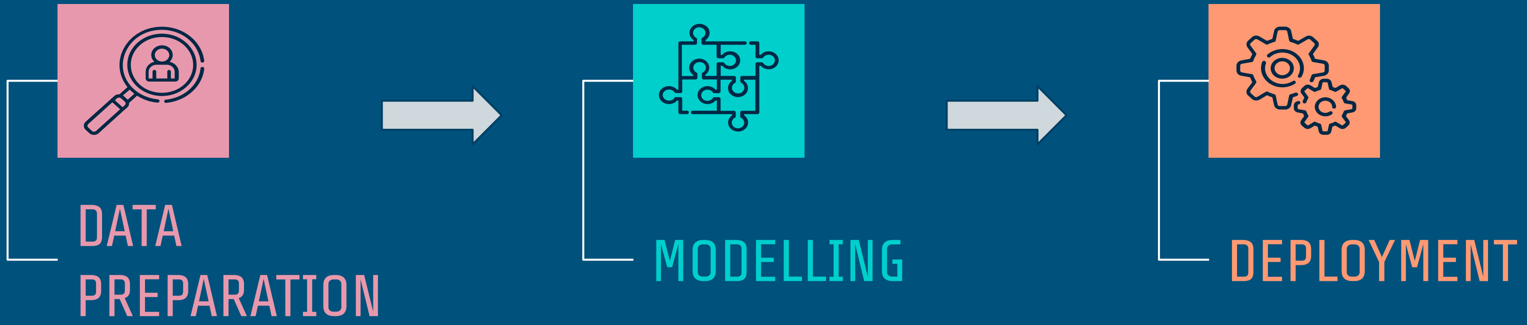
Includes nearby suburbs

Property alert Off

Sort by: Featured

New/Established: Any

METHODOLOGY/APPROACH



DATA PREPARATION

MODELLING

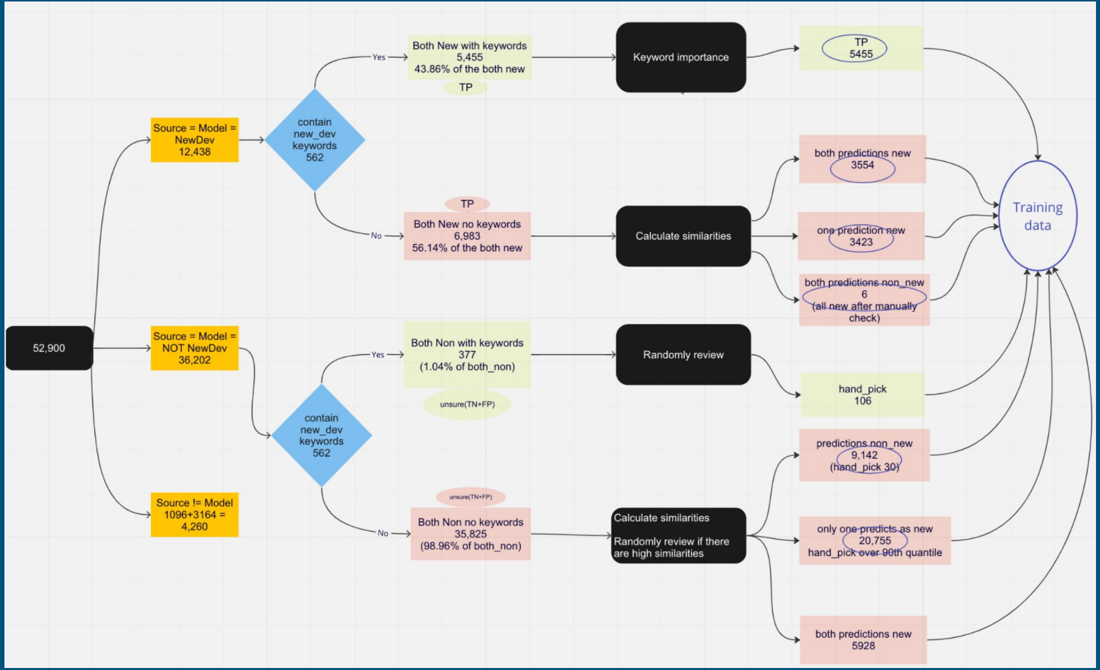
DEPLOYMENT

- Snowflake
- PyCharm
- NLP Unsupervised ML

- Vectorization/Embedding
- Feature Engineering
- NLP Supervised ML

- Docker
- CICD
- AWS ECR
- AWS S3
- Airflow
- Snowflake

PROJECT IMPLEMENTATION - DATA PREPARATION



52,900

Source = Model = NewDev 12,438

Source = Model = NOT NewDev 36,202

Source != Model 1096+3164 = 4,260

contain new_dev keywords 562

Yes → Both New with keywords 5,455 43.86% of the both new

TP

Keyword importance

TP 5455

No → Both New no keywords 6,983 56.14% of the both new

TP

Calculate similarities

both predictions new 3554

one prediction new 3423

both predictions non_new 6 (all new after manually check)

Training data

contain new_dev keywords 562

Yes → Both Non with keywords 377 (1.04% of both_non)

unsure(TN+FP)

Randomly review

hand_pick 106

No → Both Non no keywords 35,825 (98.96% of both_non)

unsure(TN+FP)

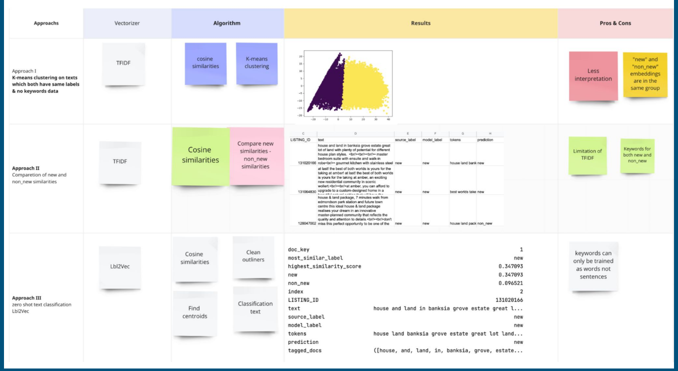
Calculate similarities

Randomly review if there are high similarities

predictions non_new 9,142 (hand_pick 30)

only one predicts as new 20,755 hand_pick over 90th quantile

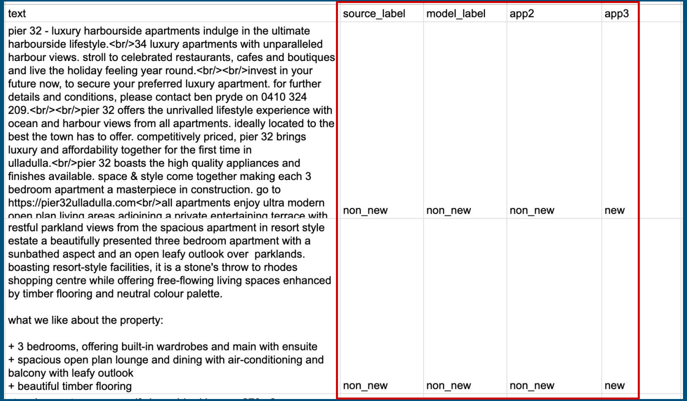
both predictions new 5928



Approachs	Vectorizer	Algorithm	Results	Pros & Cons
Approach I **K-means clustering on texts which both have same labels & no keywords data**	TFIDF	cosine similarities K-means clustering		Less interpretation "new" and "non_new" embeddings are in the same group
Approach II Comparetion of new and non_new similarities	TFIDF	Cosine similarities Compare new similarities - non_new similarities		Limitation of TFIDF Keywords for both new and non_new
Approach III zero shot text classification Lbl2Vec	Lbl2Vec	Cosine similarities Clean outliners Find centroids Classification text		keywords can only be trained as words not sentences

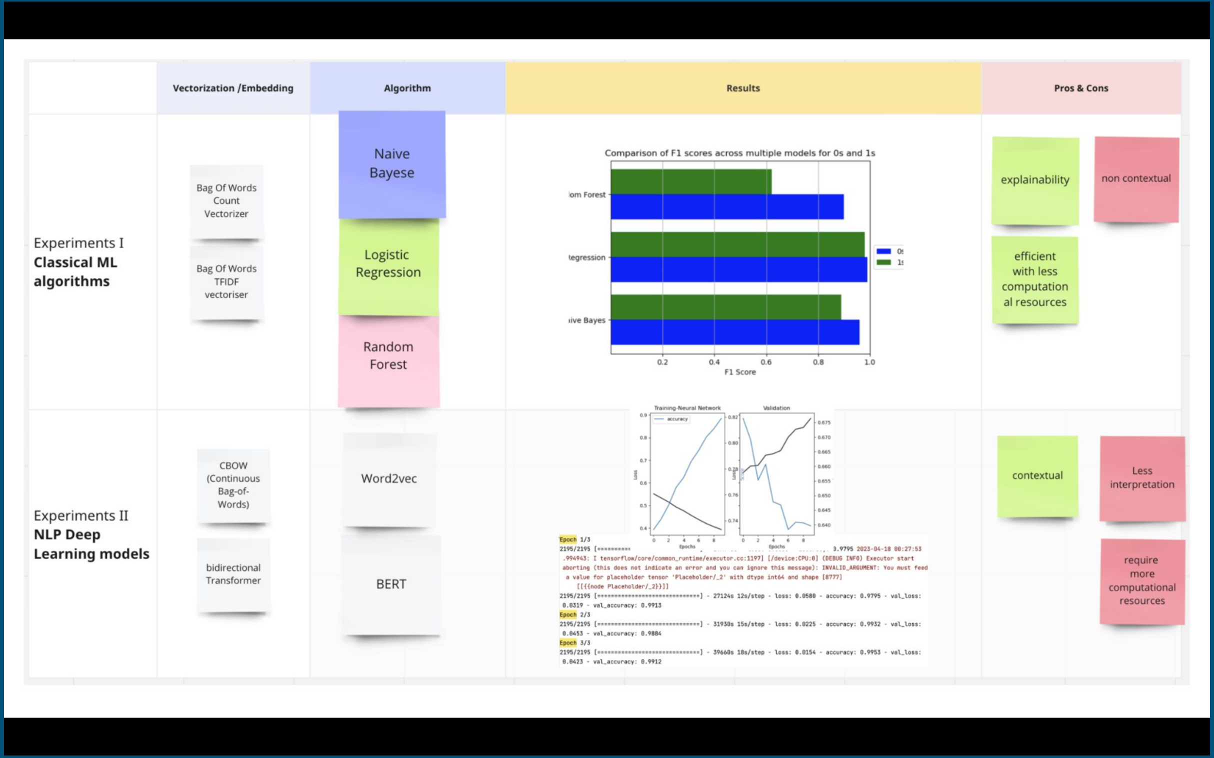
C	D	E	F	G	H
LISTING_ID	text	source_label	model_label	tokens	prediction
131020166	house and land in banksia grove estate great lot of land with plenty of potential for different house plan styles. <br/><br/><br/>• master bedroom suite with ensuite and walk-in robe<br/>• gourmet kitchen with stainless steel	new	new	house land bank	new
131064830	at last! the best of both worlds is yours for the taking at amber! at last! the best of both worlds is yours for the taking at amber, an exciting new residential community in scenic wollert.<br/><br/>at amber, you can afford to upgrade to a custom-designed home in a	new	new	best worlds take	new
128047002	house & land package, 7 minutes walk from edmondson park station and future town centre this ideal house & land package realises your dream in an innovative master-planned community that reflects the quality and attention to details.<br/><br/>don't miss this perfect opportunity to be one of the	new	new	house land pack	non_new

```
doc_key                                                       1
most_similar_label                                          new
highest_similarity_score                               0.347093
new                                                    0.347093
non_new                                                0.096521
index                                                         2
LISTING_ID                                            131020166
text                   house and land in banksia grove estate great l...
source_label                                                new
model_label                                                 new
tokens                 house land banksia grove estate great lot land...
prediction                                                  new
tagged_docs            ([house, and, land, in, banksia, grove, estate...
```



text	source_label	model_label	app2	app3
pier 32 - luxury harbourside apartments indulge in the ultimate harbourside lifestyle.<br/>34 luxury apartments with unparalleled harbour views. stroll to celebrated restaurants, cafes and boutiques and live the holiday feeling year round.<br/><br/>invest in your future now, to secure your preferred luxury apartment. for further details and conditions, please contact ben pryde on 0410 324 209.<br/><br/>pier 32 offers the unrivalled lifestyle experience with ocean and harbour views from all apartments. ideally located to the best the town has to offer. competitively priced, pier 32 brings luxury and affordability together for the first time in ulladulla.<br/>pier 32 boasts the high quality appliances and finishes available. space & style come together making each 3 bedroom apartment a masterpiece in construction. go to https://pier32ulladulla.com<br/>all apartments enjoy ultra modern open plan living areas adjoining a private entertaining terrace with	non_new	non_new	non_new	new
restful parkland views from the spacious apartment in resort style estate a beautifully presented three bedroom apartment with a sunbathed aspect and an open leafy outlook over parklands. boasting resort-style facilities, it is a stone's throw to rhodes shopping centre while offering free-flowing living spaces enhanced by timber flooring and neutral colour palette. what we like about the property: + 3 bedrooms, offering built-in wardrobes and main with ensuite + spacious open plan lounge and dining with air-conditioning and balcony with leafy outlook + beautiful timber flooring	non_new	non_new	non_new	new

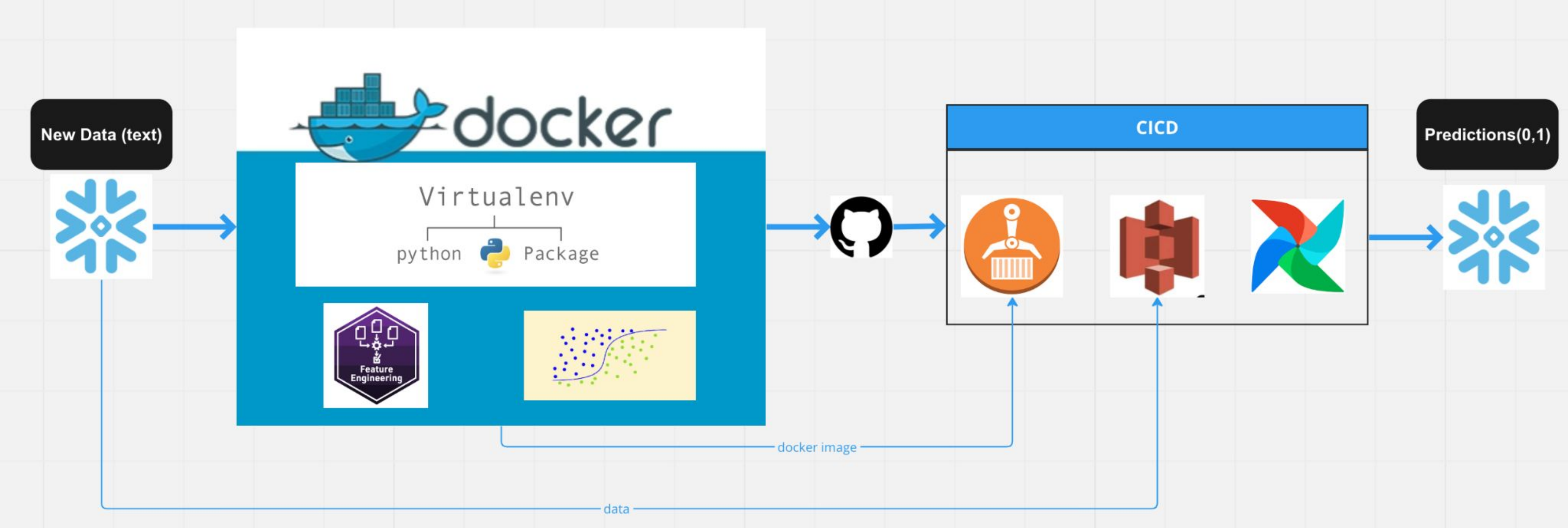
PROJECT IMPLEMENTATION – MODELLING



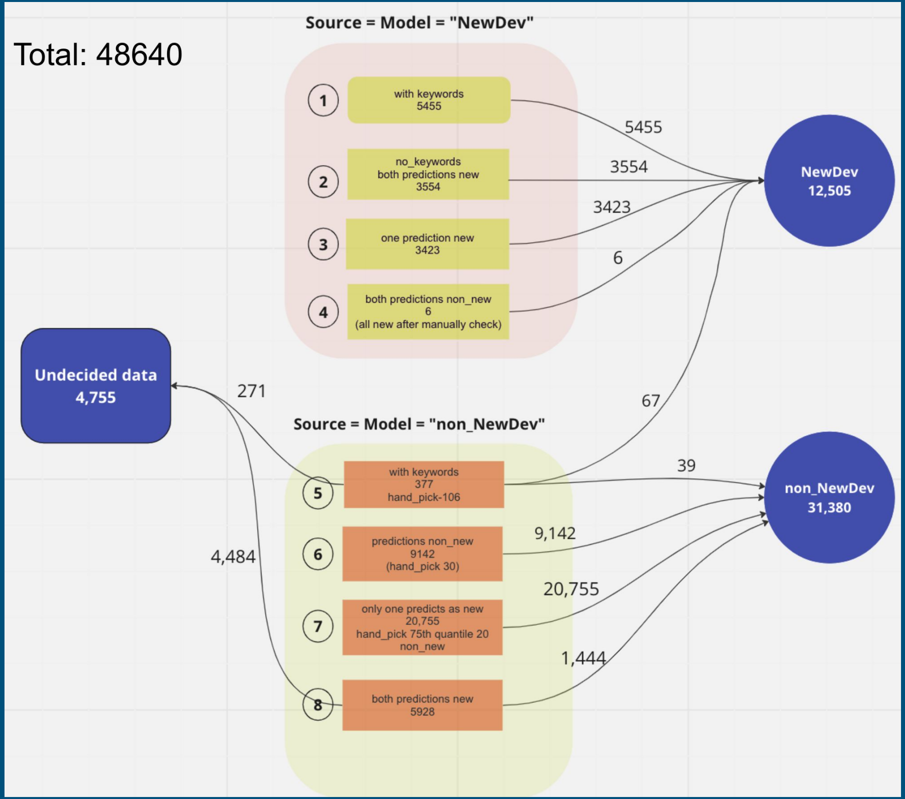
	Vectorization /Embedding	Algorithm	Results	Pros & Cons
Experiments I **Classical ML algorithms**	Bag Of Words Count Vectorizer Bag Of Words TFIDF vectoriser	Naive Bayese Logistic Regression Random Forest		explainability efficient with less computational resources non contextual
Experiments II **NLP Deep Learning models**	CBOW (Continuous Bag-of-Words) bidirectional Transformer	Word2vec BERT		contextual Less interpretation require more computational resources

```
Epoch 1/3
2195/2195 [==========                                          0.9795 2023-04-18 00:27:53
 .994943: I tensorflow/core/common_runtime/executor.cc:1197] [/device:CPU:0] (DEBUG INFO) Executor start
 aborting (this does not indicate an error and you can ignore this message): INVALID_ARGUMENT: You must feed
  a value for placeholder tensor 'Placeholder/_2' with dtype int64 and shape [8777]
     [[{{node Placeholder/_2}}]]
2195/2195 [==============================] - 27124s 12s/step - loss: 0.0580 - accuracy: 0.9795 - val_loss:
 0.0319 - val_accuracy: 0.9913
Epoch 2/3
2195/2195 [==============================] - 31930s 15s/step - loss: 0.0225 - accuracy: 0.9932 - val_loss:
 0.0453 - val_accuracy: 0.9884
Epoch 3/3
2195/2195 [==============================] - 39660s 18s/step - loss: 0.0154 - accuracy: 0.9953 - val_loss:
 0.0423 - val_accuracy: 0.9912
```

PROJECT IMPLEMENTATION – DEPLOYMENT



RESULTS/ACHIEVEMENTS – DATA PREPARATION



Total: 48640

Source = Model = "NewDev"

1 with keywords 5455

2 no_keywords both predictions new 3554

3 one prediction new 3423

4 both predictions non_new 6 (all new after manually check)

5455

3554

3423

6

NewDev 12,505

Undecided data 4,755

271

67

4,484

Source = Model = "non_NewDev"

5 with keywords 377 hand_pick-106

6 predictions non_new 9142 (hand_pick 30)

7 only one predicts as new 20,755 hand_pick 75th quantile 20 non_new

8 both predictions new 5928

39

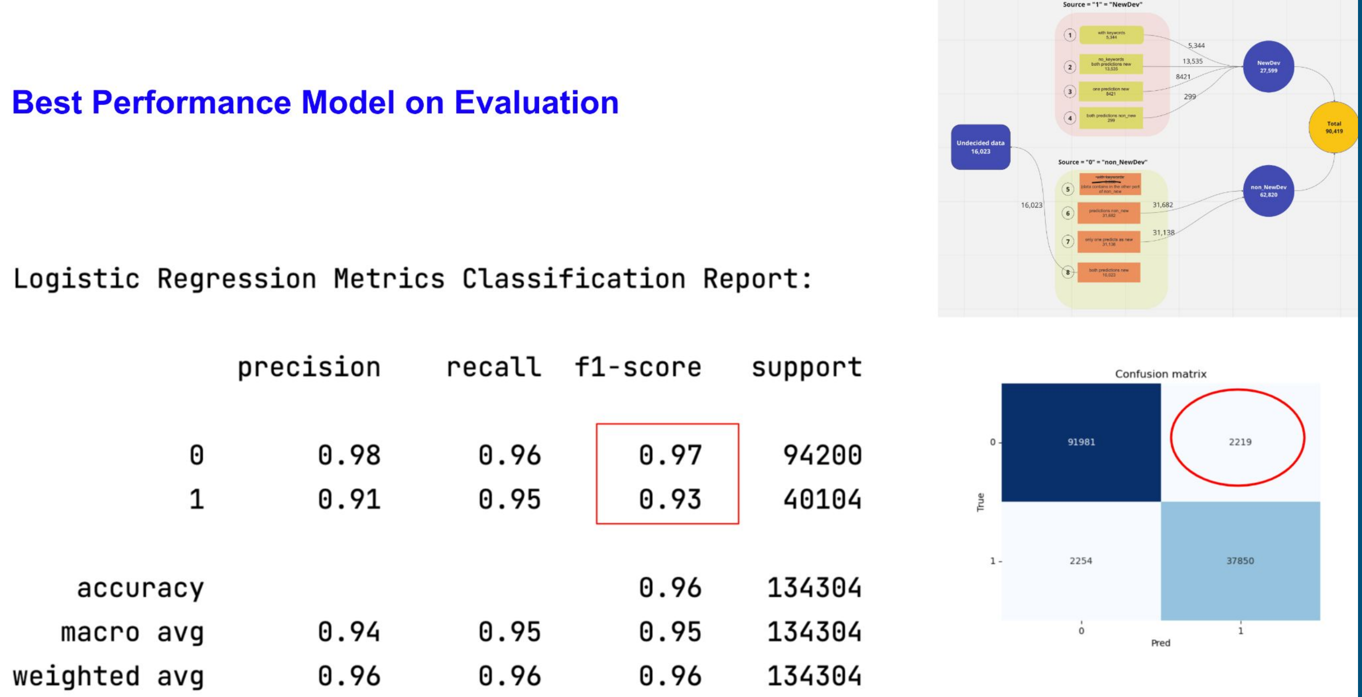
9,142

20,755

1,444

non_NewDev 31,380

RESULTS/ACHIEVEMENTS – MODELLING



Best Performance Model on Evaluation

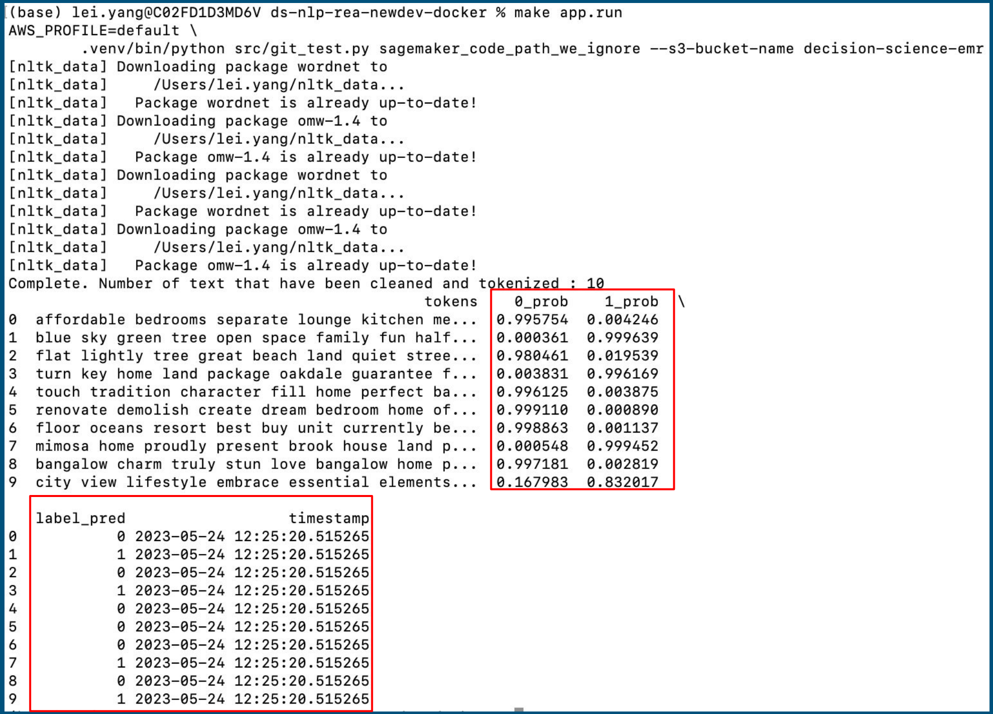
```
Logistic Regression Metrics Classification Report:

              precision    recall  f1-score   support

           0       0.98      0.96      0.97     94200
           1       0.91      0.95      0.93     40104

    accuracy                           0.96    134304
   macro avg       0.94      0.95      0.95    134304
weighted avg       0.96      0.96      0.96    134304
```

RESULTS/ACHIEVEMENTS – DEPLOYMENT



```
(base) lei.yang@C02FD1D3MD6V ds-nlp-rea-newdev-docker % make app.run
AWS_PROFILE=default \
        .venv/bin/python src/git_test.py sagemaker_code_path_we_ignore --s3-bucket-name decision-science-emr
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/lei.yang/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     /Users/lei.yang/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/lei.yang/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     /Users/lei.yang/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
Complete. Number of text that have been cleaned and tokenized : 10
                                              tokens    0_prob    1_prob  \
0  affordable bedrooms separate lounge kitchen me...  0.995754  0.004246
1  blue sky green tree open space family fun half...  0.000361  0.999639
2  flat lightly tree great beach land quiet stree...  0.980461  0.019539
3  turn key home land package oakdale guarantee f...  0.003831  0.996169
4  touch tradition character fill home perfect ba...  0.996125  0.003875
5  renovate demolish create dream bedroom home of...  0.999110  0.000890
6  floor oceans resort best buy unit currently be...  0.998863  0.001137
7  mimosa home proudly present brook house land p...  0.000548  0.999452
8  bangalow charm truly stun love bangalow home p...  0.997181  0.002819
9  city view lifestyle embrace essential elements...  0.167983  0.832017

   label_pred                  timestamp
0           0 2023-05-24 12:25:20.515265
1           1 2023-05-24 12:25:20.515265
2           0 2023-05-24 12:25:20.515265
3           1 2023-05-24 12:25:20.515265
4           0 2023-05-24 12:25:20.515265
5           0 2023-05-24 12:25:20.515265
6           0 2023-05-24 12:25:20.515265
7           1 2023-05-24 12:25:20.515265
8           0 2023-05-24 12:25:20.515265
9           1 2023-05-24 12:25:20.515265
```

CONCLUSION

LESSONS LEARNED

- Modelling is not the only aspect
- Data preparation plays crucial role

FUTURE IMPROVEMENT

- Regularly update data labels
- Understanding project purpose and overall scope from beginning

EXPERIENCE

- Real world data science project
- Complete project lifecycle and pipeline

RECOMMENDATIONS

- The descriptions of new construction properties have maintained a format that attracts potential buyers. This result can assist in refining data cleaning process and model architecture.

- Closely monitor any changes in the description format(more creative or AI-generated). In such cases, leveraging NLP deep learning models could provide an advantage.

THANKS