Project

Fake-bill Classification

Fake Invoices:

- Fake invoice:

The “Invoices” that are usually treated as ‘fake’ are those where in the GST invoices are raised by an entity without actual supply of goods or services or payment of GST.

Potential motives for using fake invoices:

> Inflating turnover for the purpose of:

a) Availing higher Credit Limit/ Overdraft from Banks
b) Obtaining bank loans
c) Improving valuations for IPO or sale of stake
d) Obtaining contracts including Government contracts

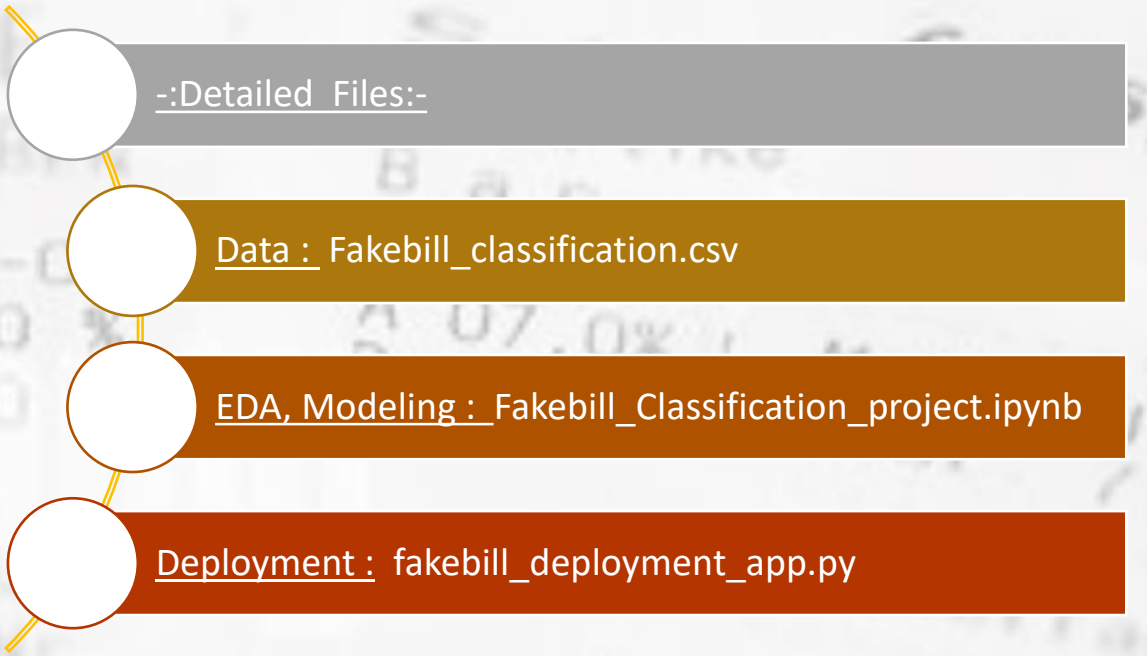
> Booking fake purchases for getting Income-tax benefits by:

a) Showing reduced profit margins and higher expenses
b) Avoiding payment of Income-tax by reducing net profit

> Cash generation/ diversion of company funds

> Laundering of money

Project Implementation Process:



Dataset: Fakebill_classification.csv

- About the Dataset

This dataset has 1500 rows and 7 columns:

1. is_genuine: Whether the bill is fake or not. (boolean)
2. diagonal: diagonal measurements in mm (float)
3. height_left: the height of the left side in mm(float)
4. height_right: the height of the right side in mm (float)
5. margin_low: the lower margin in mm (float)
6. margin_up: the upper margin in mm (float)
7. length: the total length in mm (float)

Fakebill_classification.csv

	is_genuine	diagonal	height_left	height_right	margin_low	margin_up	length
0	True	171.81	104.86	104.95	4.52	2.89	112.83
1	True	171.46	103.36	103.66	3.77	2.99	113.09
2	True	172.69	104.48	103.50	4.40	2.94	113.16
3	True	171.36	103.91	103.94	3.62	3.01	113.51
4	True	171.73	104.28	103.46	4.04	3.48	112.54
...	...	...	...	...	...	...	...
1495	False	171.75	104.38	104.17	4.42	3.09	111.28
1496	False	172.19	104.63	104.44	5.27	3.37	110.97
1497	False	171.80	104.01	104.12	5.51	3.36	111.95
1498	False	172.06	104.28	104.06	5.17	3.46	112.25
1499	False	171.47	104.15	103.82	4.63	3.37	112.07

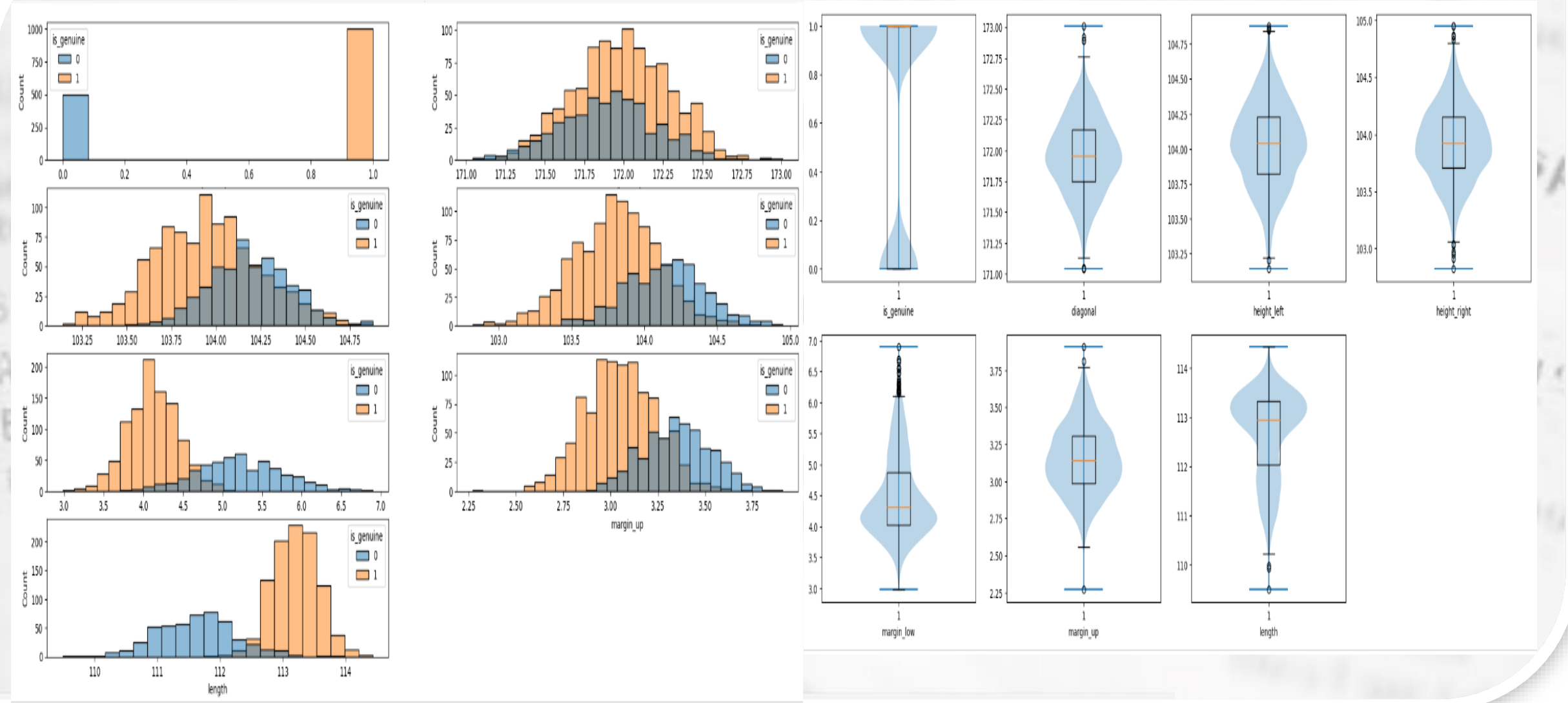
1500 rows × 7 columns

EDA Observations:

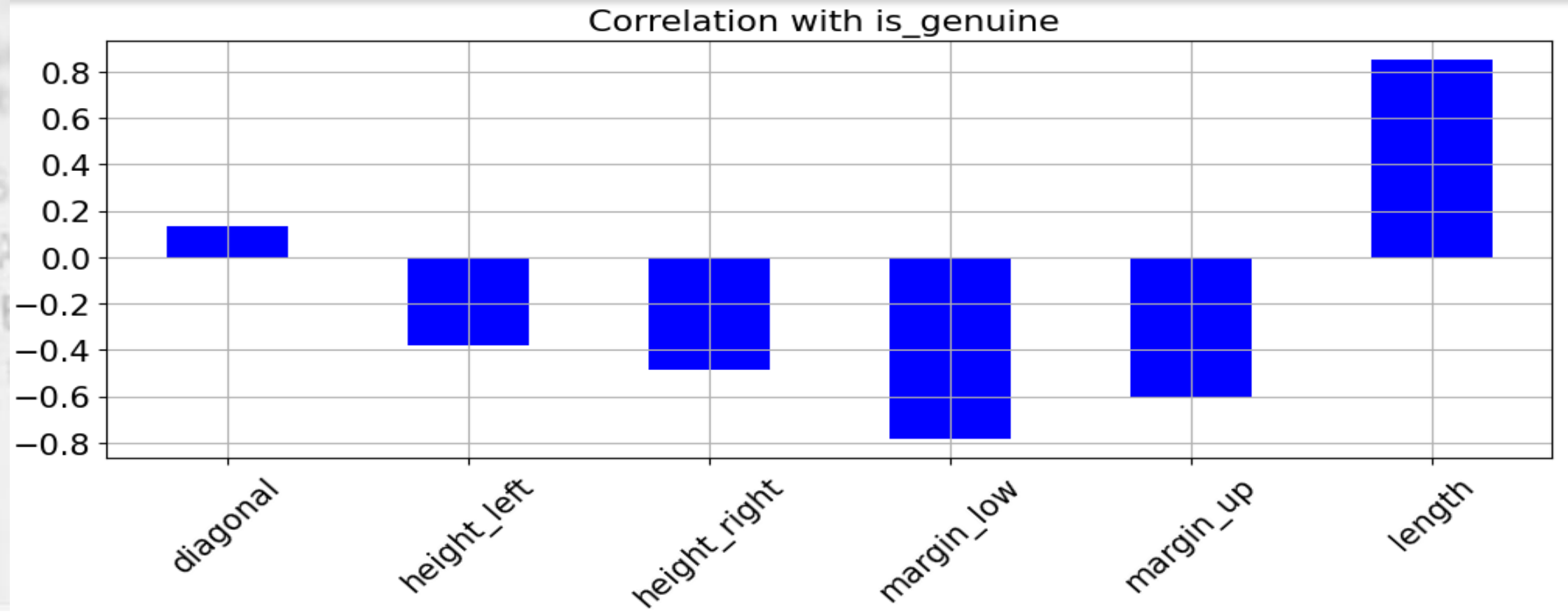
- There are 1500 rows and 7 columns in the given dataset.
- All featues have the right data type with the data values.
- There are no duplicate records in the given dataset.
- There are 37 null values in the margin_low feature. Rest All doesn't have any null values.
- There are

```
Distinct values in feature  is_genuine  are  2
Distinct values in feature  diagonal  are  159
Distinct values in feature  height_left  are  155
Distinct values in feature  height_right  are  170
Distinct values in feature  margin_low  are  285
Distinct values in feature  margin_up  are  123
Distinct values in feature  length  are  336
```

Data Visuals:



➢ Among all features, length and is_genuine are highly and positively correlated, Margin_ low and is_genuine are highly and negatively correlated.



❖Conclusion about models:

Except for Bernoulli naive Bayes and SVM without parameters, all models are giving us with around 99% accuracy.

	Model	Train score(in %) (without parameter tuning)	Test Score(in %) (without parameter tuning)	Train score(in %) (with parameter tuning)	Test Score(in %) (with parameter tuning)
0	Guassian Naive Bayes	99.33	99.24	99.33	99.24
1	Multinominal Naive Bayes	93.13	93.94	93.28	94.39
2	Bernoulli Naive Bayes	50.22	49.55	50.22	49.55
3	Complement Naive Bayes	93.28	94.39	93.28	94.39
4	Bagging Classifier	99.85	98.94	99.03	98.79
5	LogisticRegression	99.03	98.94	99.25	99.09
6	DecisionTreeClassifier	100.00	98.18	99.55	98.64
7	RandomForestClassifier	100.00	99.39	99.25	98.94
8	svm	50.22	49.55	99.40	99.24
9	KNeighborsClassifier	99.25	99.24	100.00	99.55
10	GradientBoostingClassifier	100.00	98.79	99.40	99.09
11	AdaBoostClassifier	100.00	98.94	99.70	98.79

❖Selected Model For Deployment

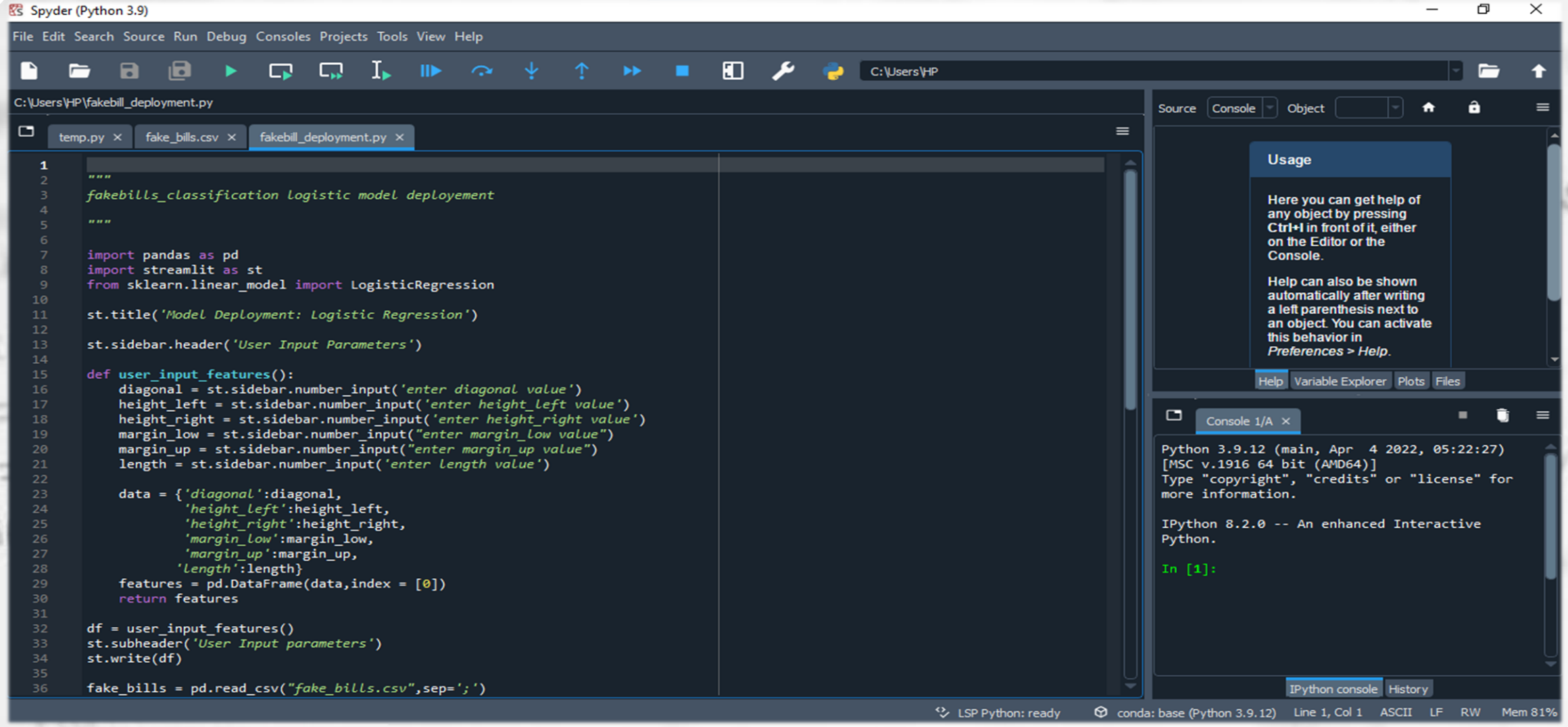
❖ Logistic Regression

❑Deployment Environment:

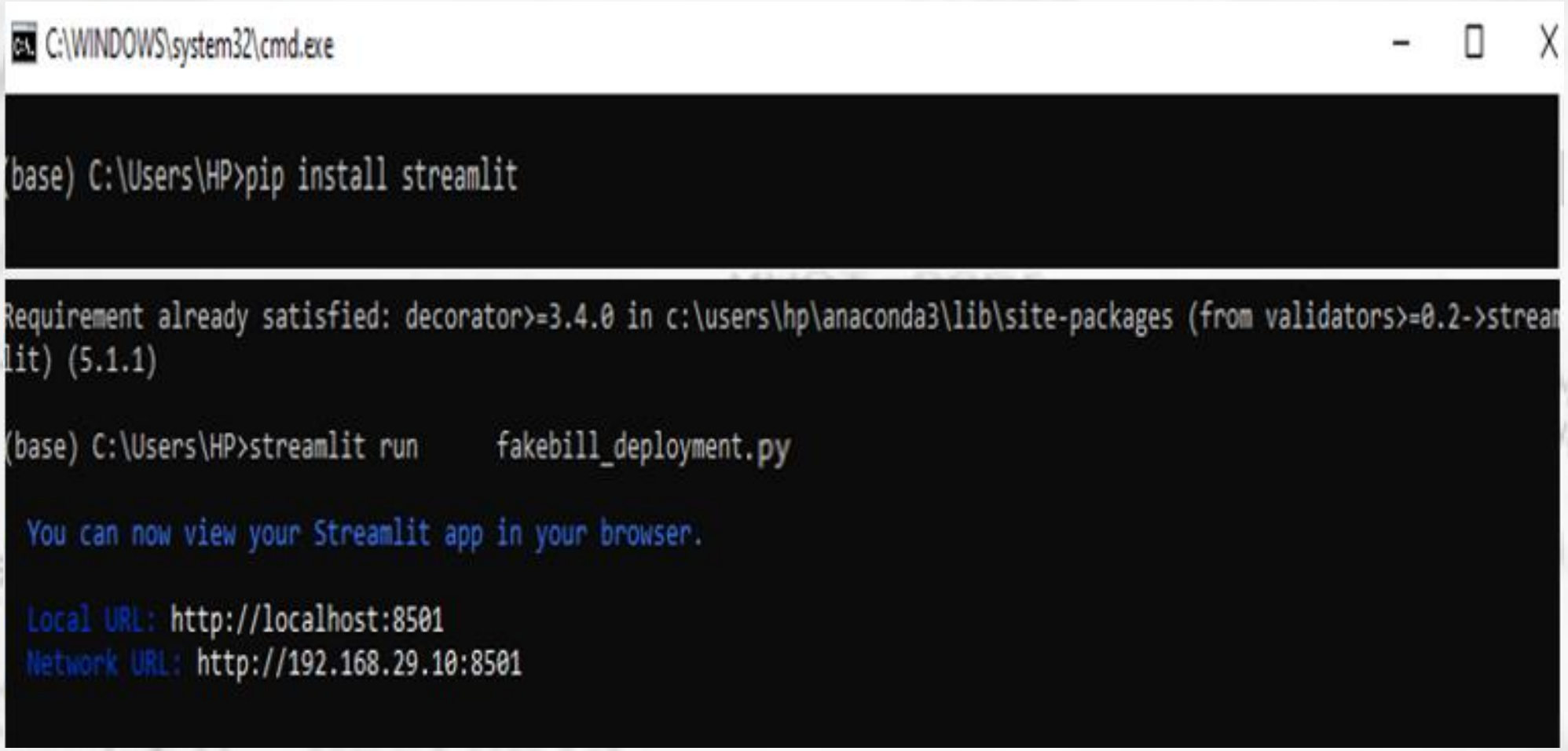
❑Streamlit

❑ Spyder

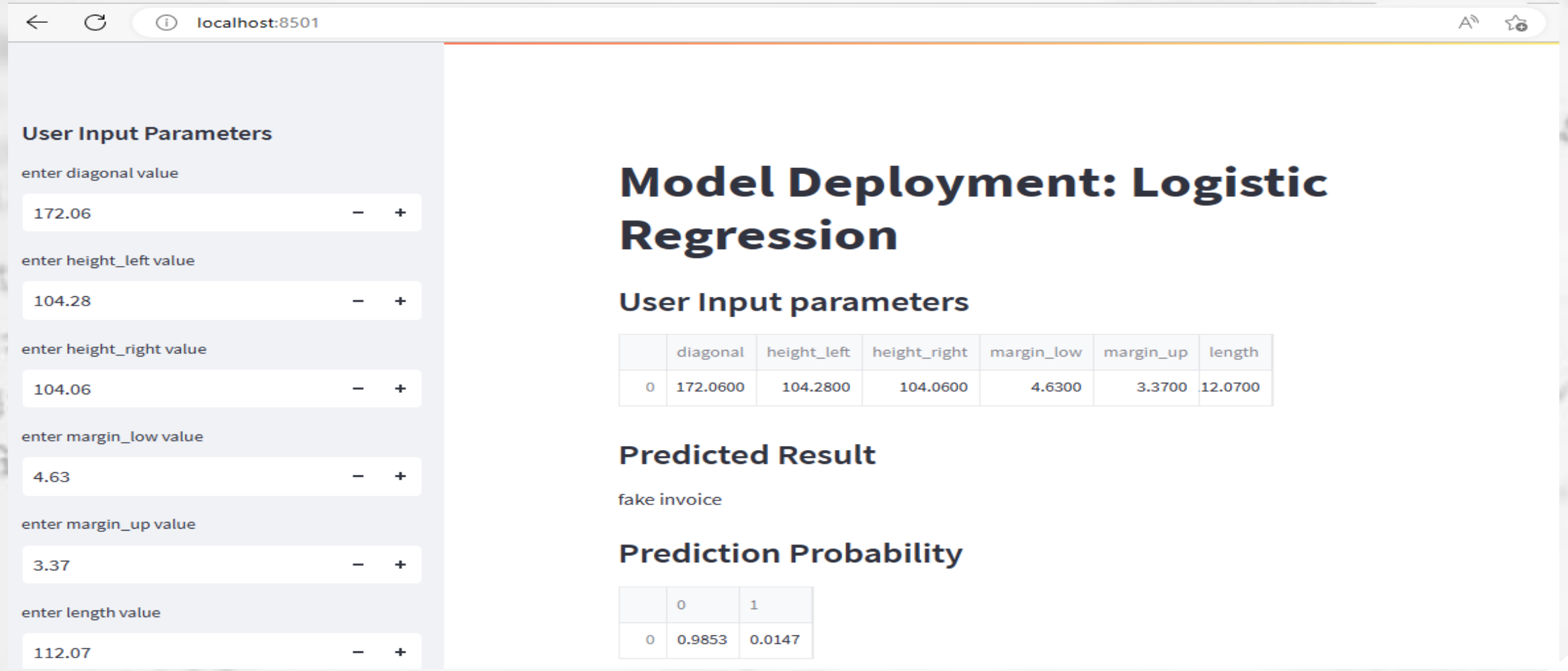
- Setting up – Code in spyder for deployment



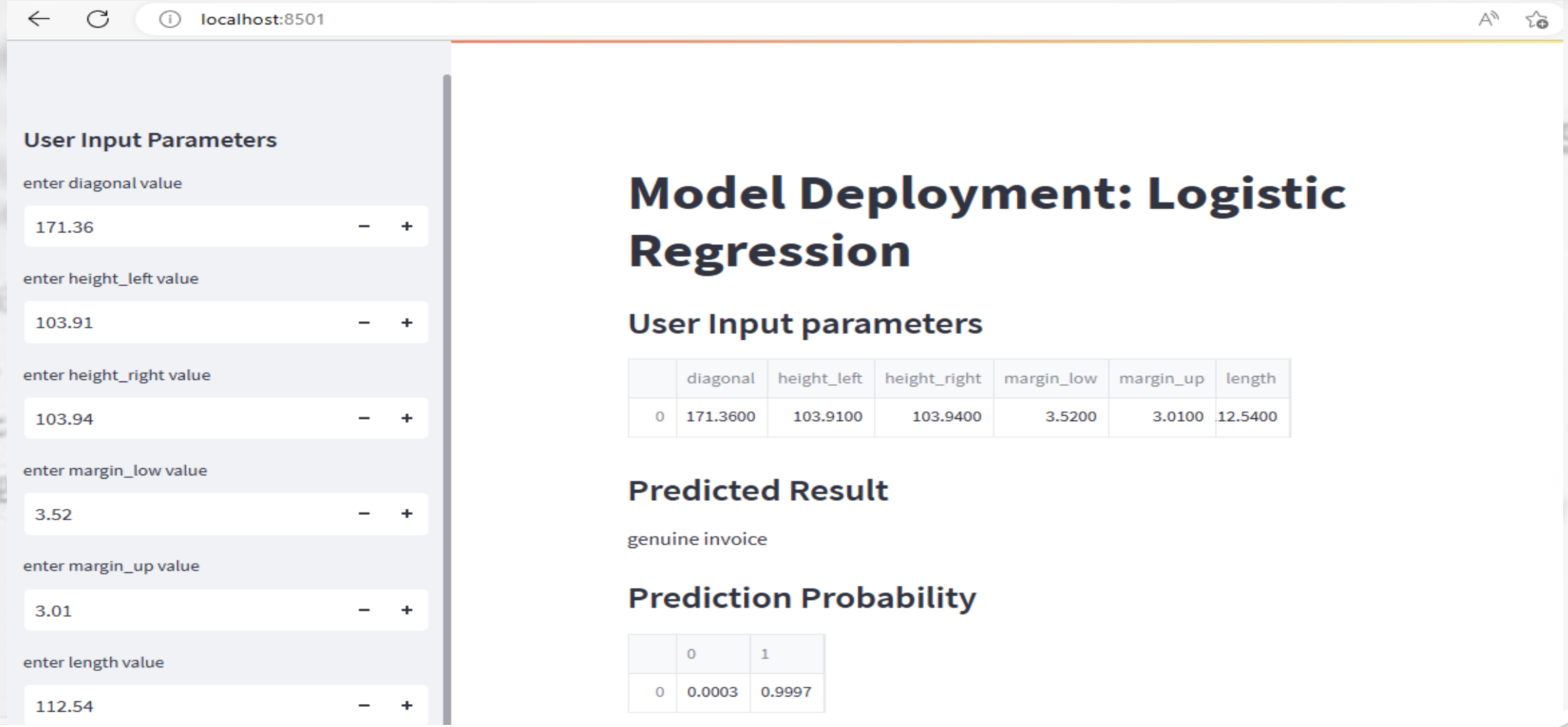
Creating Deployment Page using Streamlit:



Testing the Deployment Page: fake invoice



Testing the Deployment Page: genuine invoice



THANK YOU