MLops Assignment 2
Nitish Bhardwaj (B21AI056)

The bike sharing dataset is loaded which has columns related to calendar information, weather conditions, and sales information.

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0	0	1	1

There is a direct relationship between *casual, registered* and *cnt* column: *casual* + *registered* = *cnt.*

Therefore, *registered* and *casual* columns are dropped to reduce redundancy in the dataset.

Column *instant* which is just like the index is also dropped.

Column *dteday* is also dropped.

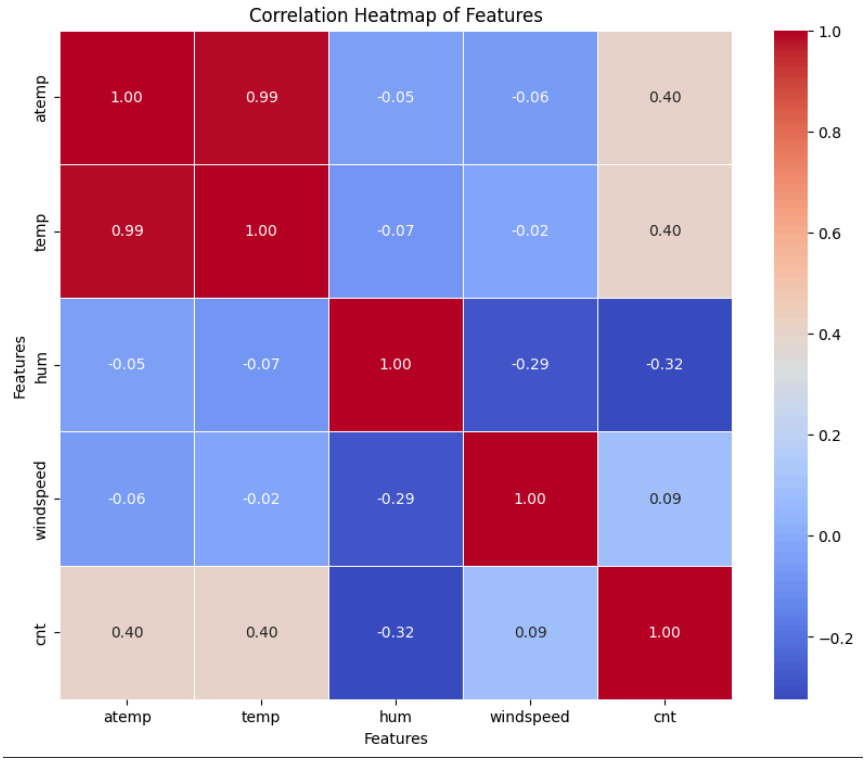
An extra categorical variable *day_night* is added based on the *hr* feature.

Categorical columns type is set to category.

We can divide the dataset into numerical columns:[*temp, atemp, hum, windspeed*] and categorical columns: [*season, yr, mnth, hr, holiday, weekday, workingday, weathersit*].

Question 1: Making two new interaction features

First heatmap is plotted among the numerical features and target (*cnt*).



We can observe that *temp* and *atemp* are highly correlated. Therefore, *atemp* is dropped to reduce redundancy in the dataset.
Remaining numerical features: *temp, hum, windspeed*.
Next, we calculated the Pearson correlation coefficient between these features and the target variable (*cnt*). To explore potential interactions, we also created new features by multiplying each pair of the existing features (e.g., *temp * hum, temp * windspeed,* etc.) and then calculated the correlation coefficients between these new features and the target.

Among these combinations, we selected those that show the highest correlation with the target variable to use in our model, as they may provide better predictive power.

	Feature_1	Feature_2	Corr(Feature_1, Target)	Corr(Feature_2, Target)	Corr(New_Feature, Target)
0	temp	hum	0.404772	-0.322911	0.072283
1	temp	windspeed	0.404772	0.093234	0.306052
2	hum	windspeed	-0.322911	0.093234	-0.058614

Hence, combinations chosen are: *temp* and *hum*, *temp* and *windspeed*.

a. *temp and hum*: Since *hum* has a negative correlation with the target, the new feature created from the interaction between *temp* and *hum* will help better capture the combined effect of these features with respect to the target. This interaction feature can provide better insights into how changes in temperature and humidity jointly affect the target variable.
b. *temp and windspeed*: *windspeed* is loosely correlated with the target. By creating a new feature that combines *temp* and *windspeed*, we can potentially capture a more meaningful relationship between these variables and the target.

Question 2: Creating pipelines for numerical and categorical variables
Numerical features are pipelined with imputation of missing values with mean of the feature and normalization using MinMax Scaler. But, for this dataset there was no missing value.

One-Hot Encoding: Each value of a categorical feature is encoded as a binary vector, with one position set to 1 and the rest to 0. To avoid multicollinearity and reduce dimensionality, the first category is dropped after encoding.

	yr	mnth	hr	holiday	weekday	workingday	temp	hum	windspeed	temp_hum	temp_windspeed	season_2	season_3	season_4	weathersit_2	weathersit_3	weathersit_4	day_night_night
0	0	1	0	0	6	0	0.224490	0.81	0.0	0.312039	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0	1	1	0	6	0	0.204082	0.80	0.0	0.282504	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
2	0	1	2	0	6	0	0.204082	0.80	0.0	0.282504	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	0	1	3	0	6	0	0.224490	0.75	0.0	0.288925	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
4	0	1	4	0	6	0	0.224490	0.75	0.0	0.288925	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Target Encoding: Categorical values are encoded based on the mean of the target variable for each category. For each unique value in the categorical feature, the corresponding target mean is calculated and used as the encoded value.

	yr	mnth	hr	holiday	weekday	workingday	temp	hum	windspeed	temp_hum	temp_windspeed	season	weathersit	day_night
0	0	1	0	0	6	0	0.224490	0.81	0.0	0.312039	0.0	111.114569	204.869272	98.894138
1	0	1	1	0	6	0	0.204082	0.80	0.0	0.282504	0.0	111.114569	204.869272	98.894138
2	0	1	2	0	6	0	0.204082	0.80	0.0	0.282504	0.0	111.114569	204.869272	98.894138
3	0	1	3	0	6	0	0.224490	0.75	0.0	0.288925	0.0	111.114569	204.869272	98.894138
4	0	1	4	0	6	0	0.224490	0.75	0.0	0.288925	0.0	111.114569	204.869272	98.894138

Question 3: Train Linear Regression model

a. **Using package:** Model is trained using the linear regression model from scikit-learn.

b. **From Scratch:** Implemented the linear regression model using the normal equation for parameters estimation.

Training:

$$\boldsymbol{\beta} = (A^T A)^{-1} A^T y$$

where $\boldsymbol{\beta}$ is the weights and bias,
A is the input features stacked with extra bias column,
and y is the target variable.

Testing:

$$y_{pred} = X_{test}.\, w + b$$

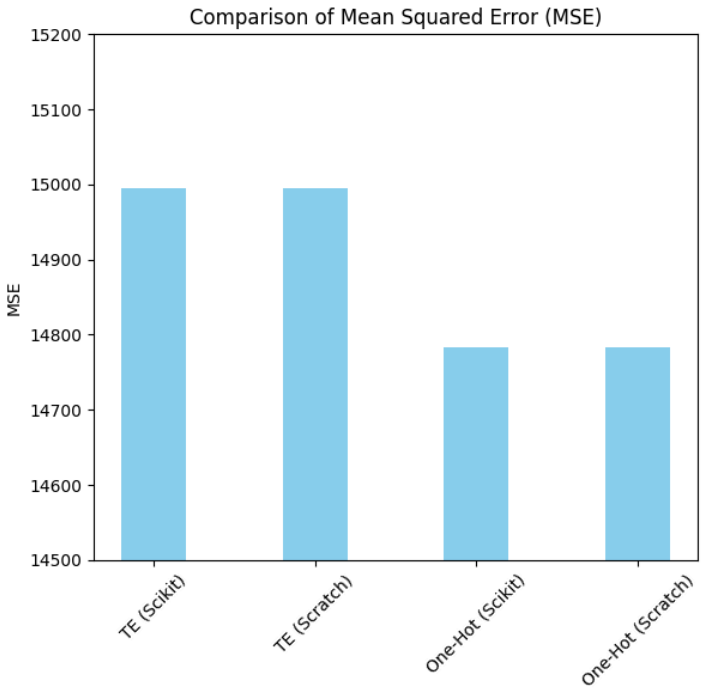
where y_{pred} is the predicted target value,
X_{test} is the input features whose target is to be calculated,
w and b are the estimated weights and biases.

Performance comparison:

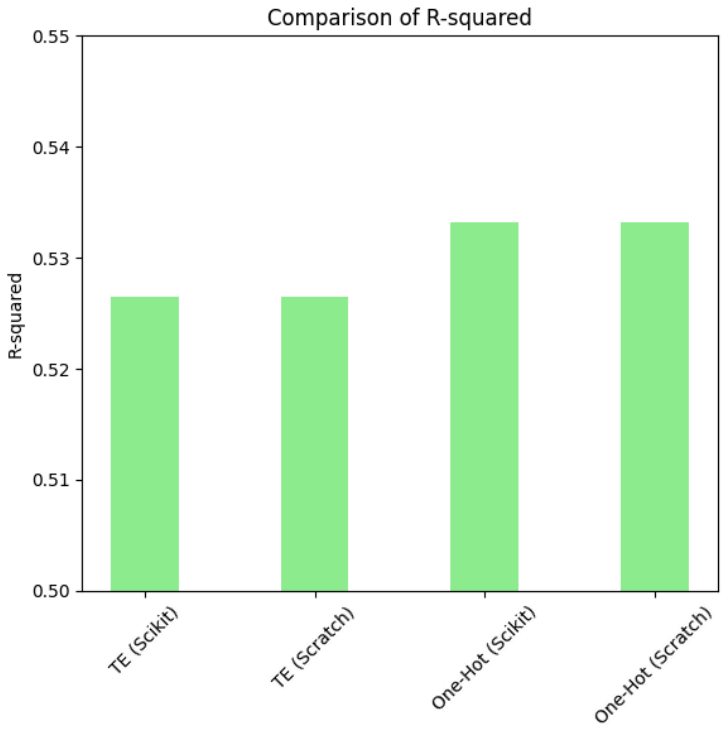
We compared the performance model implemented from scratch and that from the package. Result is that both showed same performance wrt to MSE and R2 score.
This also shows that for this dataset since it is small, the package based model also used the normal equation method.
Had the dataset been large, then we should have used an iterative method based on gradient descent approach. Also the normal equation method should not be used if $A^T A$ is not invertible, else it will lead to the singularity issue.
Since we had taken care of redundancy in the dataset, we didn't face this problem.

Along with this, we compared results with one-hot encoding and target encoding method. Result is that the model performed better with One-hot encoding.

MSE Results: It shows that when input features had categorical values encoded using one-hot encoding, the model was able to capture the underlying pattern better, hence MSE is less with one-hot encoding than Target Encoding.



R2 Score Results: A higher R2 score indicates that the model's predictions explain a greater proportion of the variance in the target variable. The R2 score was higher when using one-hot encoding for categorical features compared to Target Encoding. This means that one-hot encoding better captured the contribution of input features to the target variable.

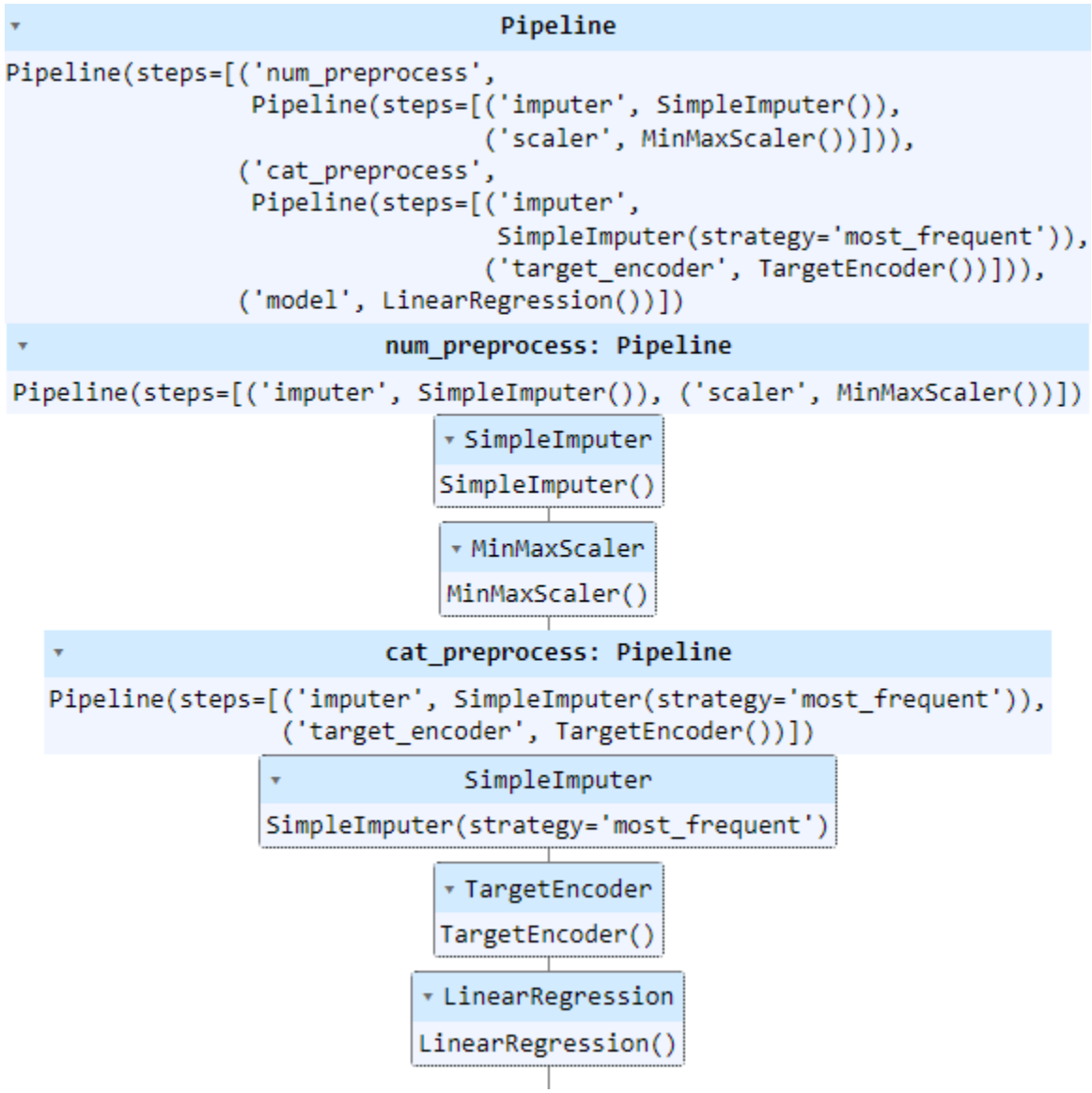


Hence, the overall model performed better with one-hot encoding over target encoding, and performance with the package model and scratch model was the same.

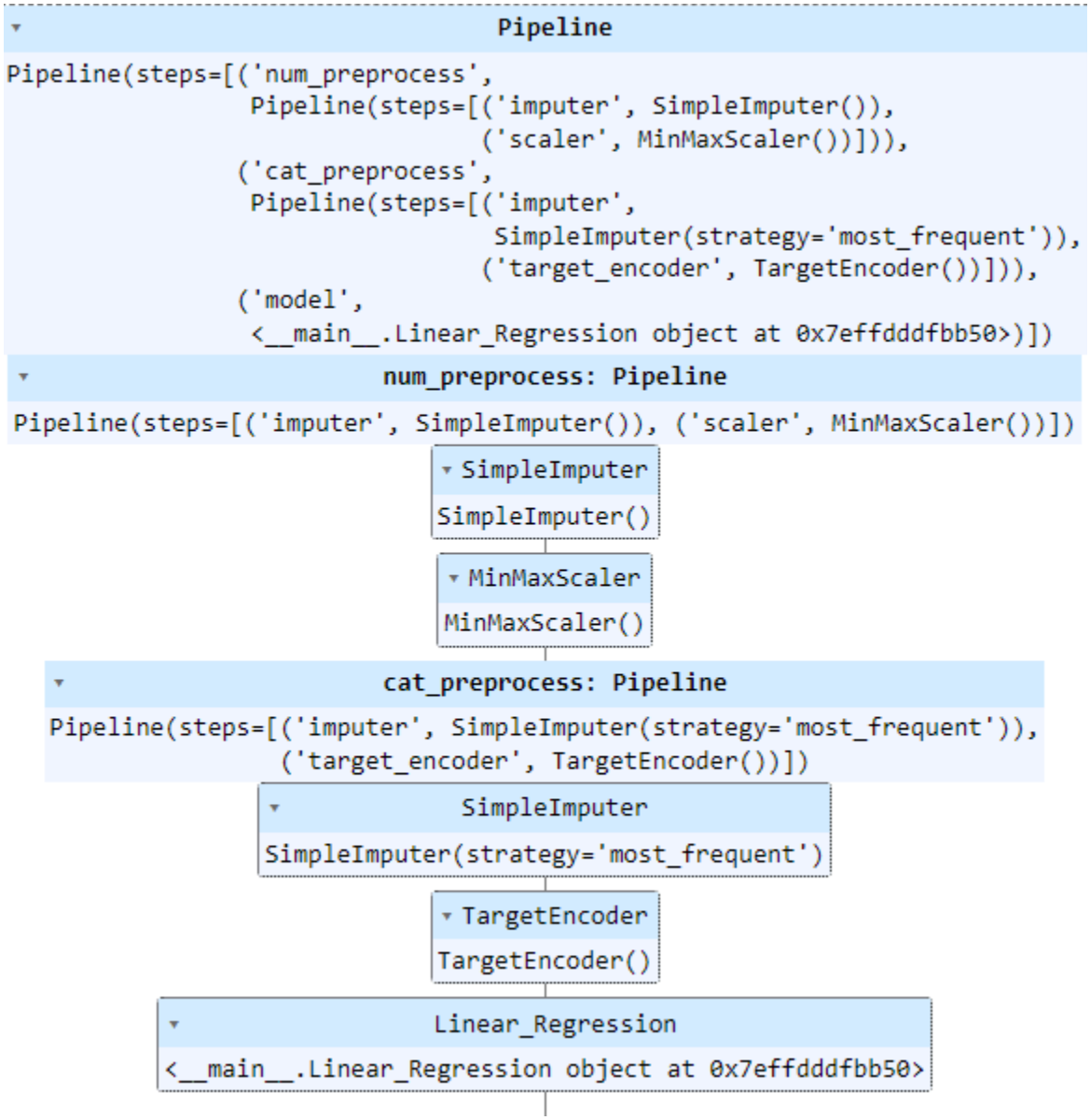
Question 4: 4. Integrating MLflow into the pipeline

ML flow is integrated with 3 pipelines, numerical, categorical then model.

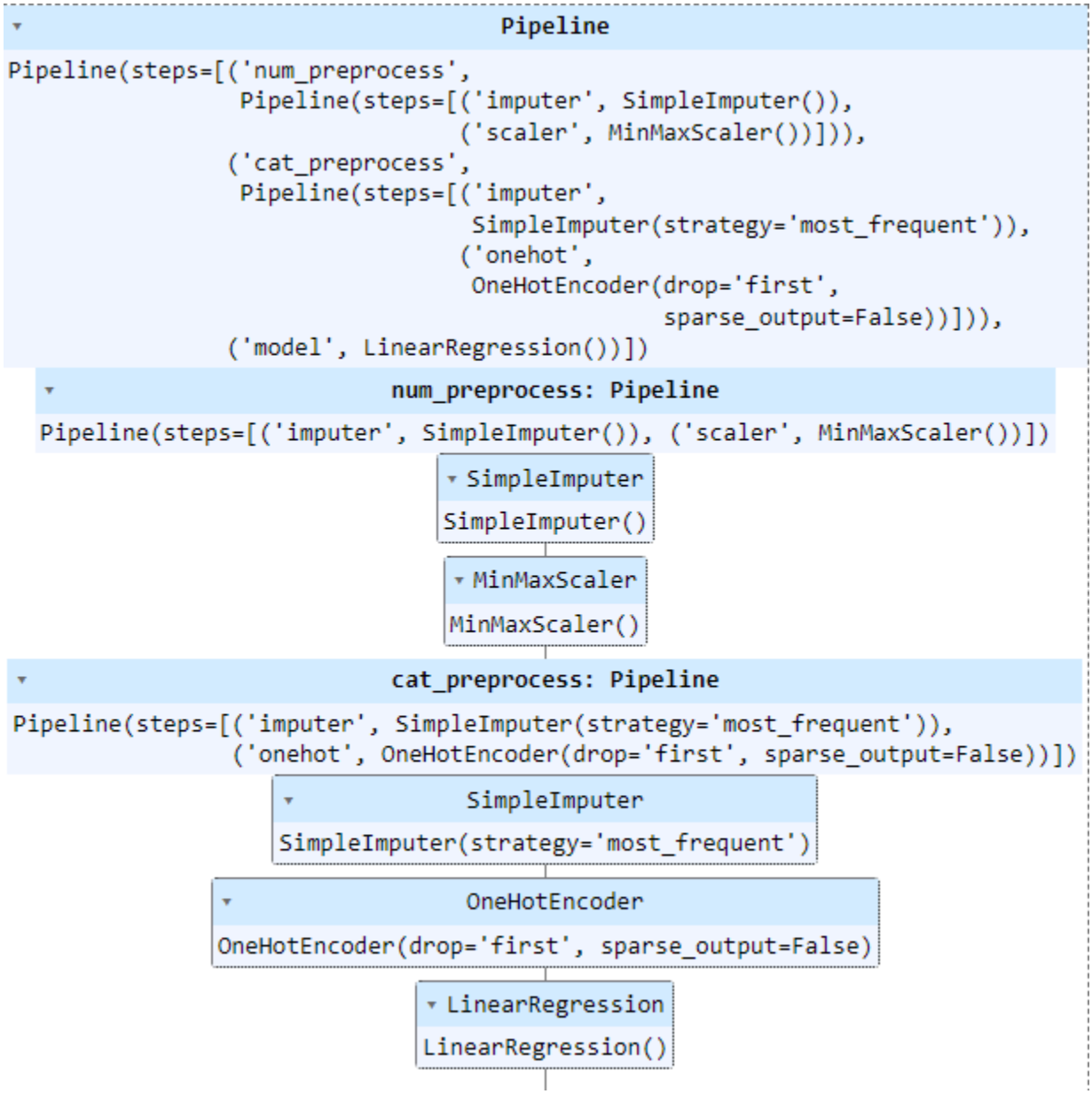
a. **Model - Package, Categorical Encoding - TargetEncoding**



b. Model -Scratch, Categorical Encoding - TargetEncoding



c. Model - Package, Categorical Encoding - OneHotEncoding



d. Model -Scratch, Categorical Encoding - OneHotEncoding