LearnX Sales Forecasting

Anushree Tomar

28-03-2020

Problem statement

LearnX is an online learning platform aimed at professionals and students. LearnX serves as a market place that allows instructors to build online courses on topics of their expertise which is later published after due diligence by the LearnX team. The platform covers a wide variety of topics including Development, Business, Finance & Accounting & Software Marketing and so on.

Effective forecasting for course sales gives essential insight into upcoming cash flow meaning business can more accurately plan the budget to pay instructors and other operational costs and invest in the expansion of the business.

Sales data for more than 2 years from 600 courses of LearnX's top domains is available along with information on:-

- Competition in the market for each course
- Course Type (Course/Program/Degree)
- Holiday Information for each day
- User Traffic on Course Page for each day

Your task is to predict the course sales for each course in the test set for the next 60 days.

Data Dictionary

The *Train data* (Historical Sales Data) has following attributes:-

Variable	Definition
ID	Unique Identifier for a row
Day_No	Day Number
Course_ID	Unique ID for a course
Course_Domain	Course Domain (Development, Finance etc.)
Course_Type	Course/Program/Degree
Short_Promotio	Whether Short Term Promotion is Live
Public_Holiday	Regional/Public Holiday
Long_Promotion	Whether Long Term Promotion is Live for the course
User_Traffic	Number of customers landing on the course page
Competition_Metric	A metric defining the strength of competition
Sales (Target)	Total Course Sales

The *Test data* (Next 60 Days)

This file contains the store and day number for which the participant needs to submit predictions/forecasts

Variable	Definition
ID	Unique Identifier for a row
Day_No	Day Number
Course_ID	Unique ID for a course
Course_Domain	Course Domain (Development, Finance etc.)
Course_Type	Course/Program/Degree
Short_Promotion	Whether Short Term Promotion is Live
Public_Holiday	Regional/Public Holiday
Long_Promotion	Whether Long Term Promotion is Live for the course
Competition_Metric	A metric defining the strength of competition

Sample Submission

This file contains the exact submission format for the forecasts. Please submit csv file only.

Variable	Definition
ID	Unique Identifier for a row
Sales (Target)	Total Course Sales predicted from the test set

Evaluation Metric

The evaluation metric for this competition is 1000*RMSLE where RMSLE is Root of Mean Squared Logarithmic Error across all entries in the test set.

Data Exploration

Now let's explore the train data:-

Train data

Top and Bottom of the data

```
##    ID Day_No Course_ID Course_Domain Course_Type Short_Promotion Public_Holiday
## 1:  1      1         1   Development      Course               0              1
## 2:  2      2         1   Development      Course               0              0
## 3:  3      3         1   Development      Course               0              0
## 4:  4      4         1   Development      Course               0              0
## 5:  5      5         1   Development      Course               0              0
## 6:  6      6         1   Development      Course               0              0
##    Long_Promotion User_Traffic Competition_Metric Sales
## 1:              1        11004              0.007    81
## 2:              1        13650              0.007    79
## 3:              1        11655              0.007    75
```

```
## 4:              1        12054              0.007    80
## 5:              1         6804              0.007    41
## 6:              1        10395              0.007    62
```

```
##        ID Day_No Course_ID      Course_Domain Course_Type Short_Promotion
## 1: 548022    877       600 Software Marketing     Program               0
## 2: 548023    878       600 Software Marketing     Program               0
## 3: 548024    879       600 Software Marketing     Program               0
## 4: 548025    880       600 Software Marketing     Program               0
## 5: 548026    881       600 Software Marketing     Program               0
## 6: 548027    882       600 Software Marketing     Program               1
##    Public_Holiday Long_Promotion User_Traffic Competition_Metric Sales
## 1:              0              1         9072               0.07   111
## 2:              0              1         8904               0.07   114
## 3:              0              1        10542               0.07   145
## 4:              0              1        13671               0.07   167
## 5:              0              1         8904               0.07   107
## 6:              0              1        11445               0.07   152
```

Change the data type of attributes.

Basic stats of the train data

```
##        ID             Day_No         Course_ID
##  Min.   :     1   2      :   600   1      :   882
##  1st Qu.:136963   3      :   600   2      :   882
##  Median :273984   4      :   600   3      :   882
##  Mean   :274007   5      :   600   4      :   882
##  3rd Qu.:411066   6      :   600   5      :   882
##  Max.   :548027   7      :   600   6      :   882
##                   (Other):508487   (Other):506795
##               Course_Domain     Course_Type     Short_Promotion Public_Holiday
##  Business            :  4410   Course :262747   0:317369        0:495885
##  Development         :264295   Degree :  1764   1:194718        1: 16202
##  Finance & Accounting: 77210   Program:247576
##  Software Marketing  :166172
##
##
##
##  Long_Promotion  User_Traffic    Competition_Metric     Sales
##  0:261693       Min.   :   168   Min.   :0.0000     Min.   :  0.0
##  1:250394       1st Qu.: 10584   1st Qu.:0.0100     1st Qu.: 84.0
##                 Median : 13776   Median :0.0350     Median :111.0
##                 Mean   : 15375   Mean   :0.0733     Mean   :120.8
##                 3rd Qu.: 18123   3rd Qu.:0.0940     3rd Qu.:146.0
##                 Max.   :100002   Max.   :0.7680     Max.   :682.0
##                                  NA's   :1764
```

courses with Zero Sales

```
##         ID Day_No Course_ID Course_Domain Course_Type Short_Promotion
```

```
##  1: 119717    363       132   Development      Course               0
##  2: 120081    727       132   Development      Course               0
##  3: 120085    731       132   Development      Course               0
##  4: 145151    731       159   Development     Program               0
##  5: 243951    371       267   Development      Course               1
##  6: 244316    736       267   Development      Course               1
##  7: 316334    360       346   Development      Course               0
##  8: 316337    363       346   Development      Course               0
##  9: 316340    366       346   Development      Course               0
## 10: 341179    691       373   Development      Course               0
## 11: 355560    736       389   Development      Course               1
## 12: 398563    731       437   Development      Course               0
## 13: 424961    731       466   Development      Course               0
## 14: 507353    736       556   Development      Course               1
##     Public_Holiday Long_Promotion User_Traffic Competition_Metric Sales
##  1:              0              0         1050              0.021     0
##  2:              0              0         1050              0.021     0
##  3:              1              0         1029              0.021     0
##  4:              1              1          756              0.004     0
##  5:              1              0          714              0.449     0
##  6:              1              0          630              0.449     0
##  7:              1              1          672              0.004     0
##  8:              0              1          588              0.004     0
##  9:              1              1          588              0.004     0
## 10:              0              0          168              0.269     0
## 11:              1              1          693              0.247     0
## 12:              1              0         1113              0.008     0
## 13:              1              0          945              0.364     0
## 14:              1              0          987              0.021     0
```

There are some courses with Development Domain having zero sales on different Day No.

Basic stats of the test data

```
##        ID              Day_No        Course_ID                   Course_Domain
##  Min.   :   883   883    :  600   1      :   60   Business            :  300
##  1st Qu.:137730   884    :  600   2      :   60   Development         :18480
##  Median :274762   885    :  600   3      :   60   Finance & Accounting: 5340
##  Mean   :274566   886    :  600   4      :   60   Software Marketing  :11880
##  3rd Qu.:410873   887    :  600   5      :   60
##  Max.   :548087   888    :  600   6      :   60
##                   (Other):32400   (Other):35640
##   Course_Type    Short_Promotion Public_Holiday Long_Promotion
##  Course :18600   0:21600         0:35605        0:17940
##  Degree :  120   1:14400         1:  395        1:18060
##  Program:17280
##
##
##
##
##  Competition_Metric
##  Min.   :0.00000
```

```
##  1st Qu.:0.01000
##  Median :0.03450
##  Mean   :0.07294
##  3rd Qu.:0.09400
##  Max.   :0.76800
##  NA's   :120
```

Checking Missing data

```
##      rows columns discrete_columns continuous_columns all_missing_columns
## 1: 512087      11               7                 4                   0
##    total_missing_values complete_rows total_observations memory_usage
## 1:                 1764        510323            5632957     26730600
```

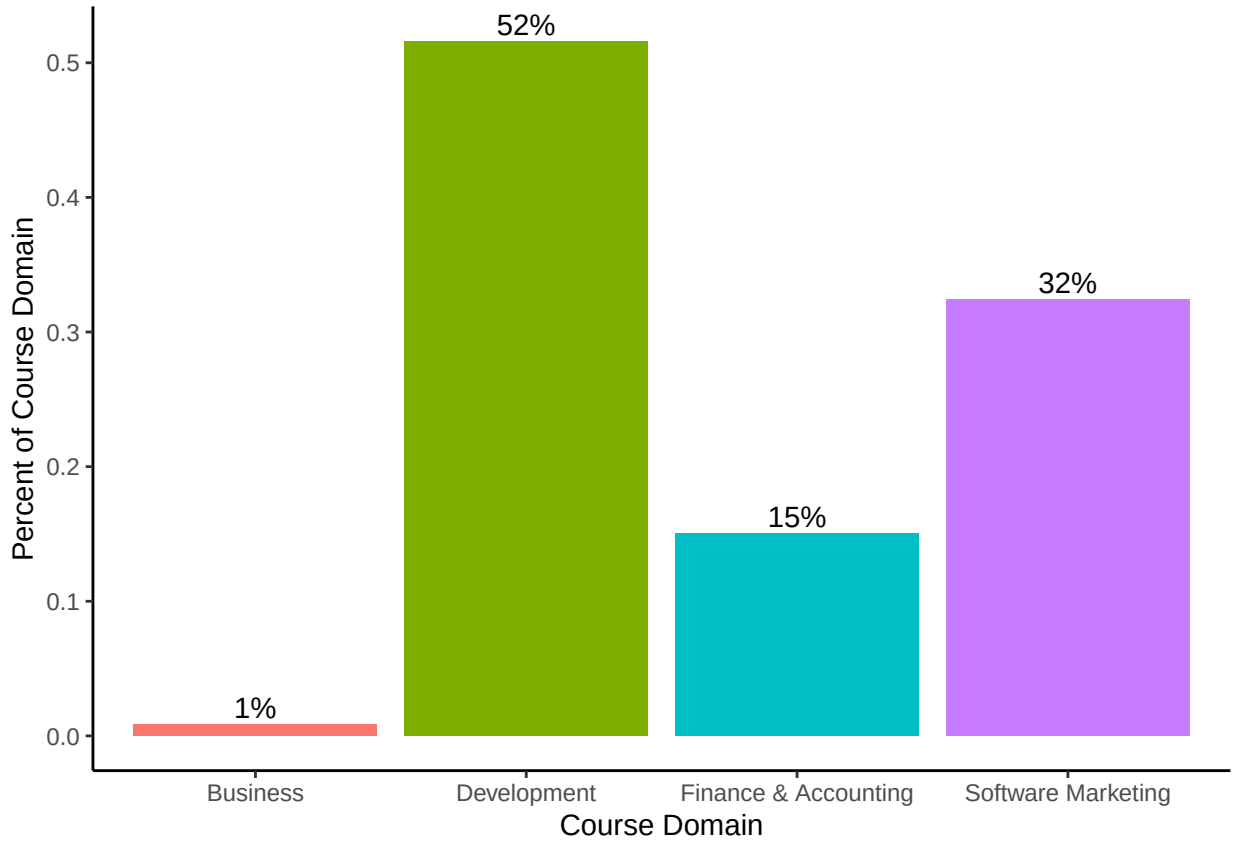
Total Sales by each Course

```
##      Course_ID Total_sales
##   1:       158       43836
##   2:       124       44274
##   3:       466       44869
##   4:       153       45128
##   5:       570       45685
##  ---
## 596:       397      241397
## 597:       304      244040
## 598:       424      256287
## 599:       225      269970
## 600:       151      297807
```

Visualization of the data

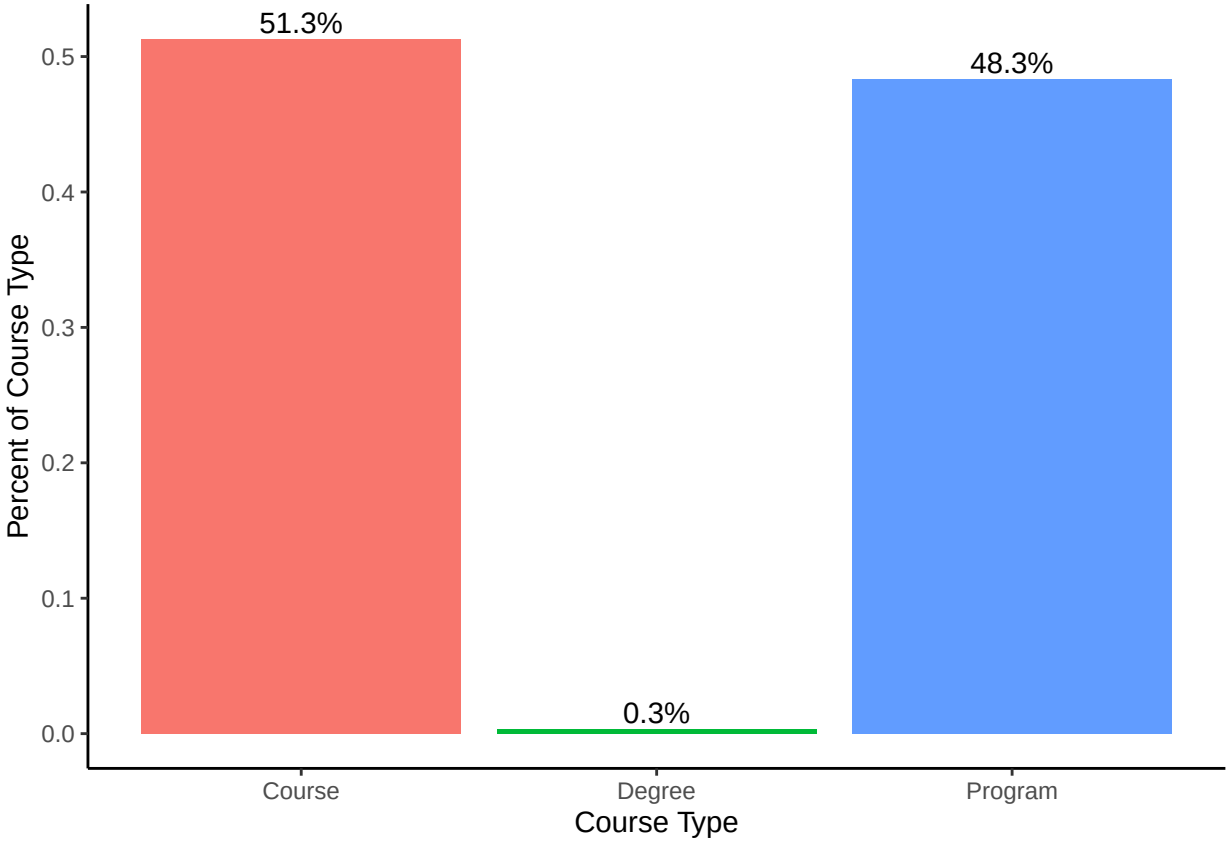
Univariate and Bivariate Analysis

Distribution of Course Domain



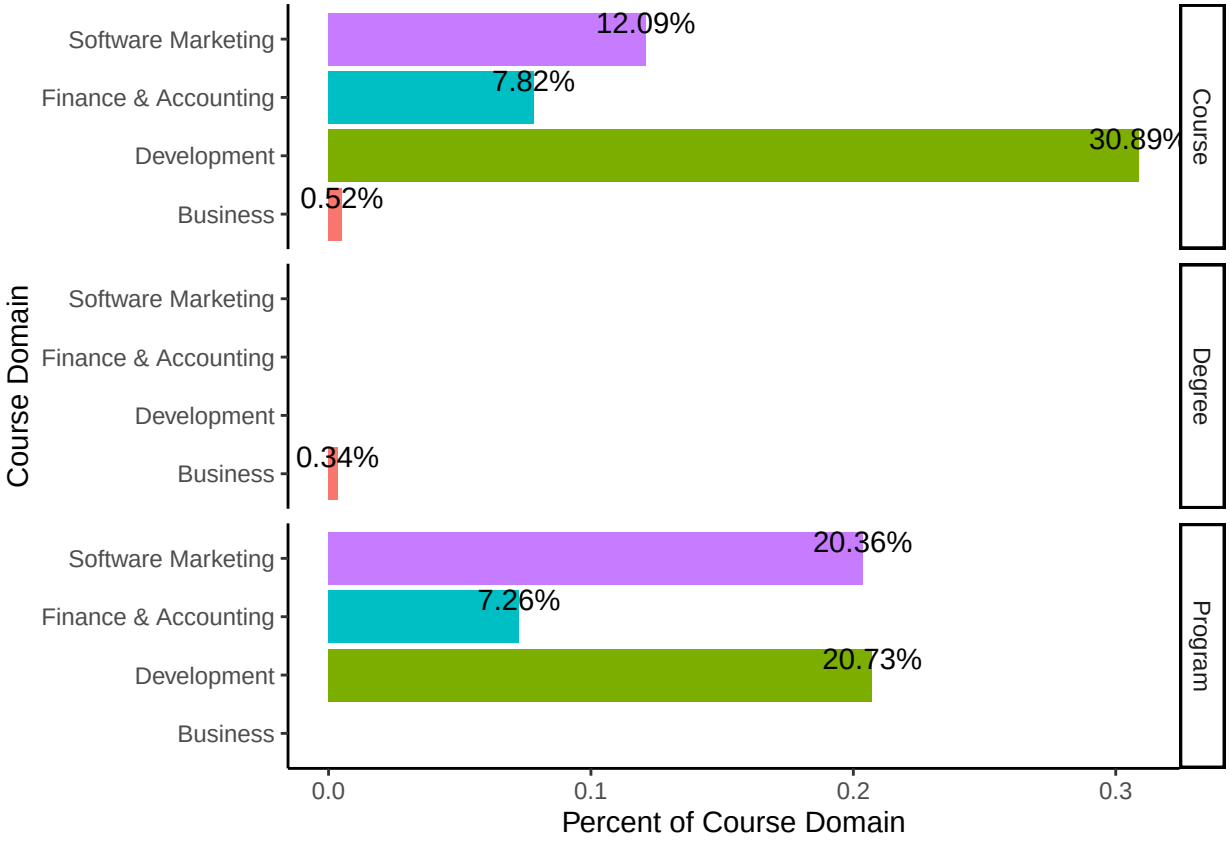
From the above graph we can see that out of 600 courses, *52%* of couses belongs to *Develpment* domain,*32%* of course belong to *Software Marketting* and *15%* of courses belongs to *Finance Accounting* while only *1%* of course belongs to *Business Domain*.

Distribution of Course Type

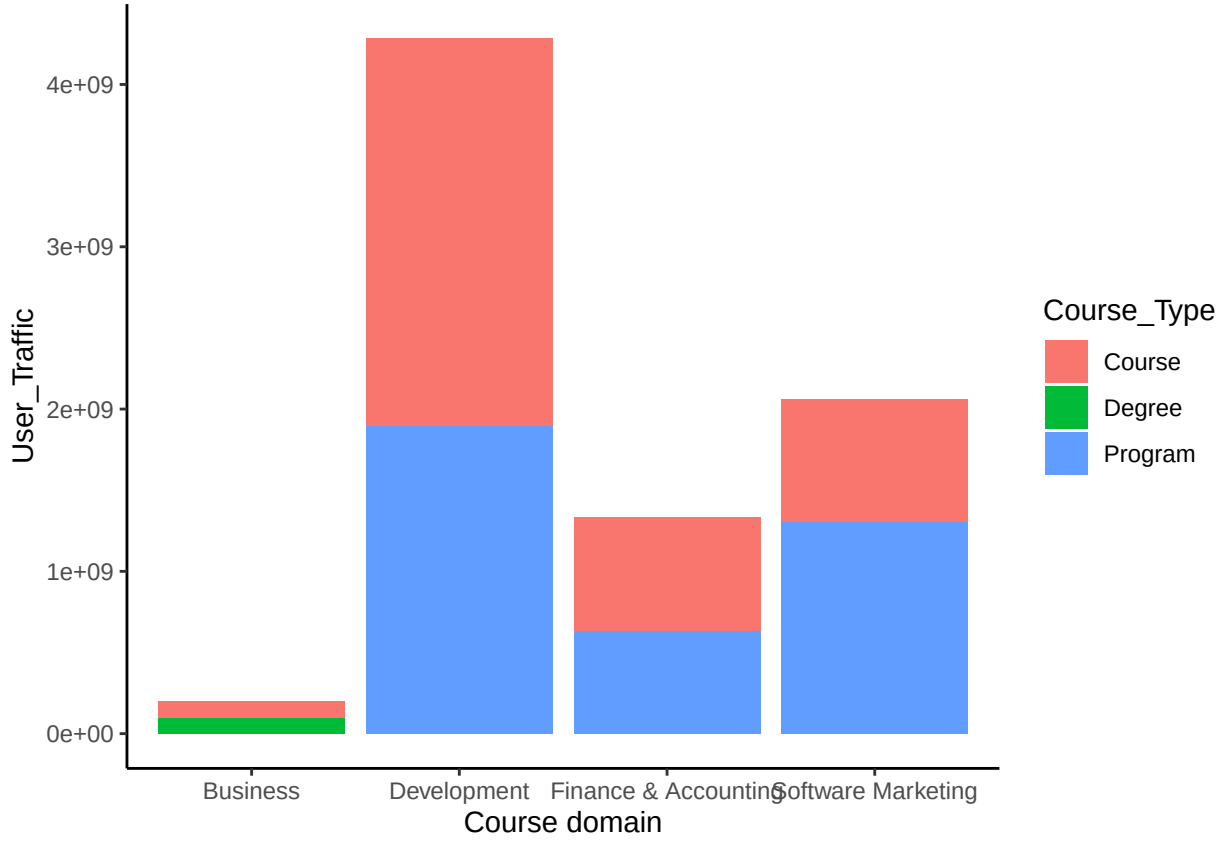


There are *51.3%* of Courses are *Course Type* and *48.3%* of courses *Progrme Type* while only *0.3%* of courses are of *Degree Type*.

Comparative analysis of Course Domain and Course Type



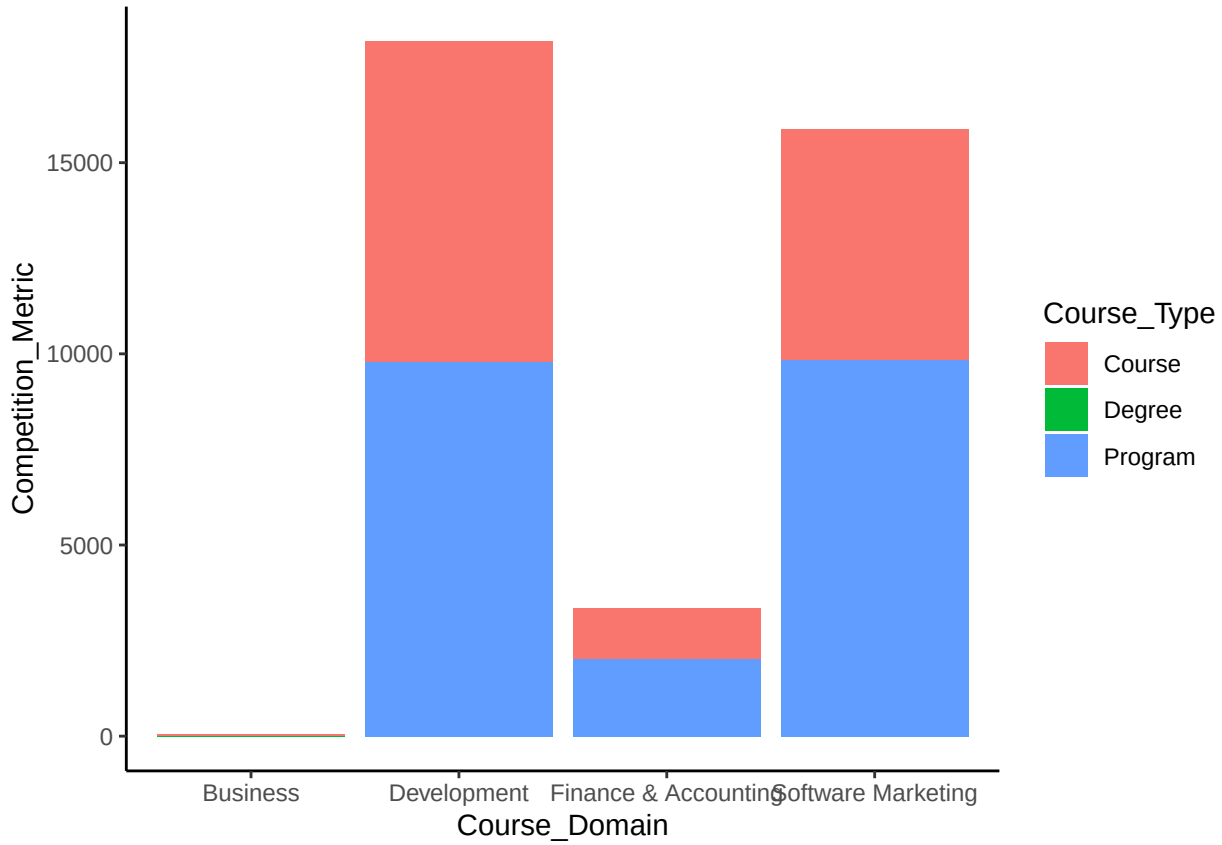
Comparative analysis of Course Domain and User traffic



From the above graph we can conclude that there are more number of *user traffic* for *Development Domain* and for *Course Type* followed by *Software Marketting* for *Program Type*.

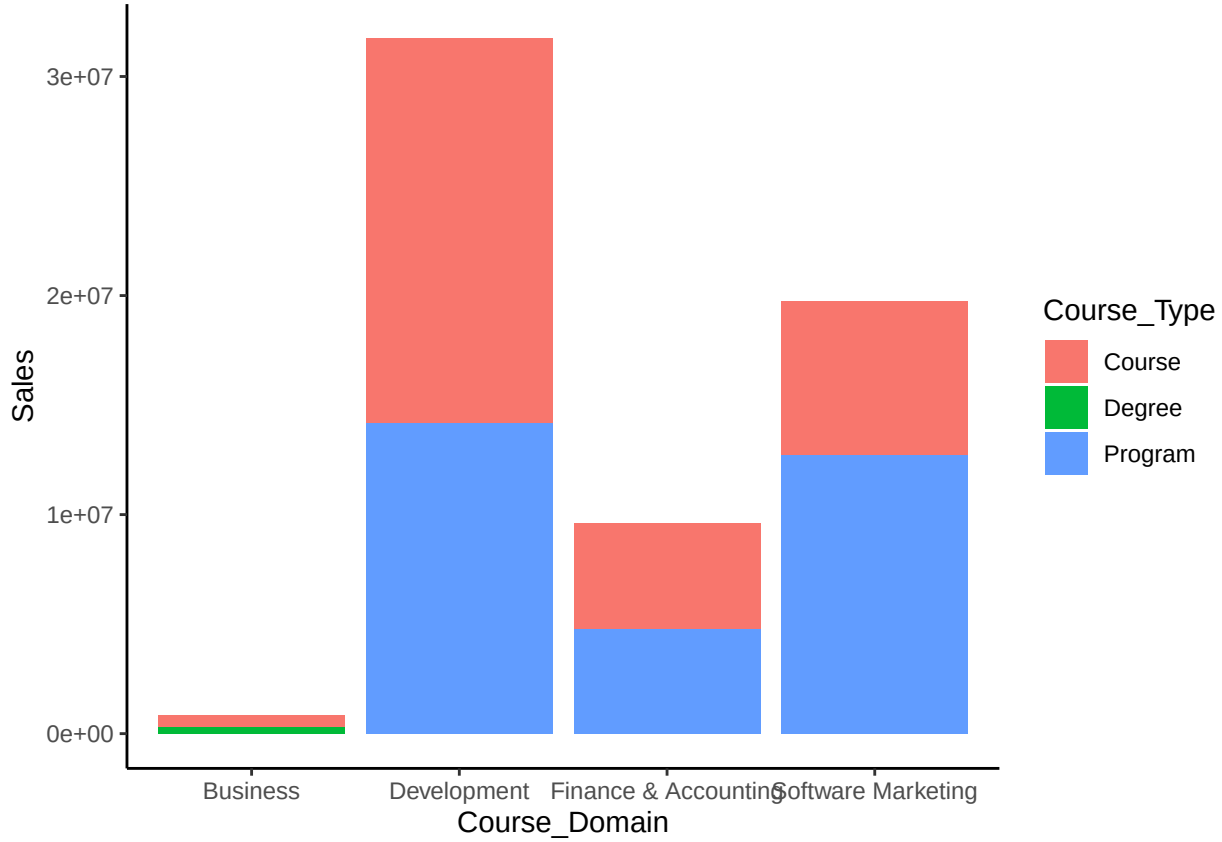
Comparative analysis of Course Domain and Competitive Metric

```
## Warning: Removed 1764 rows containing missing values (position_stack).
```



The strength of competition is more for *Development Domain* and of *Program Type*.

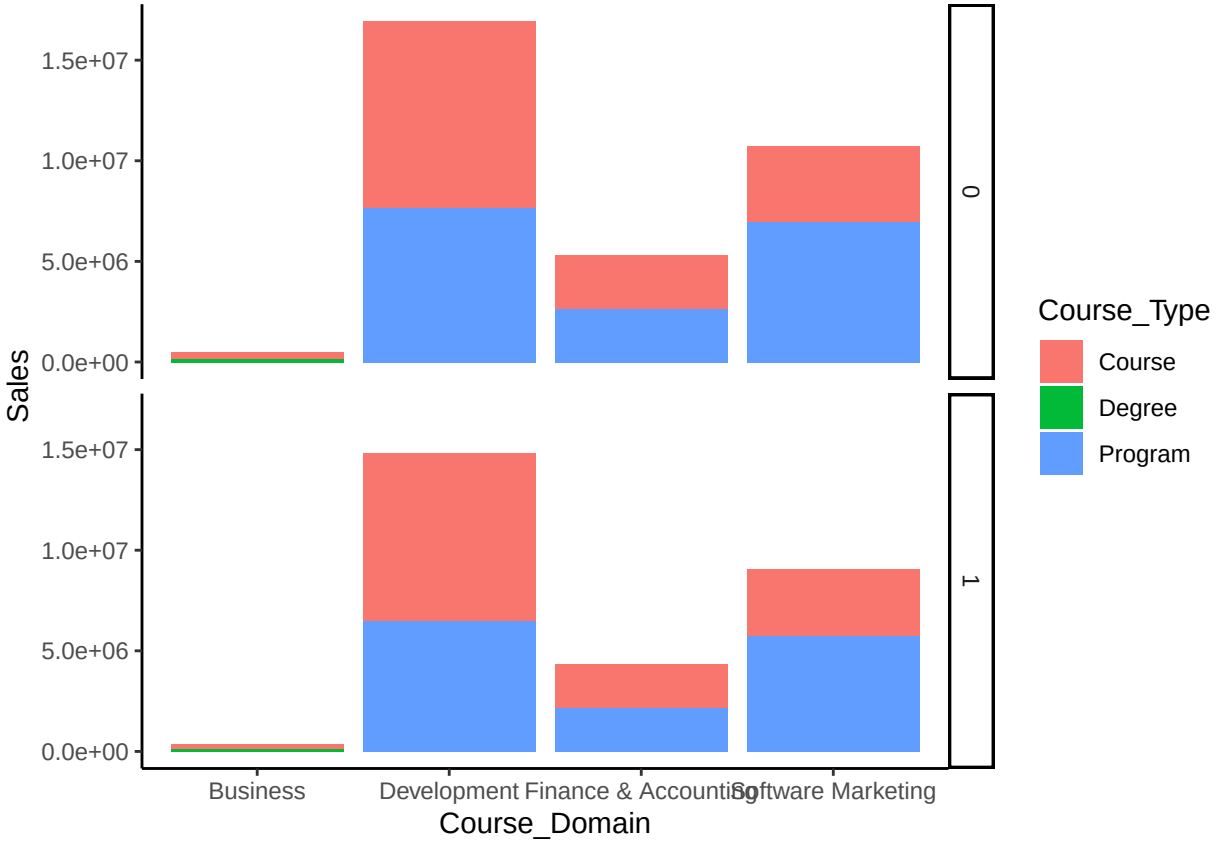
Comparative analysis of Course Domain and Sales



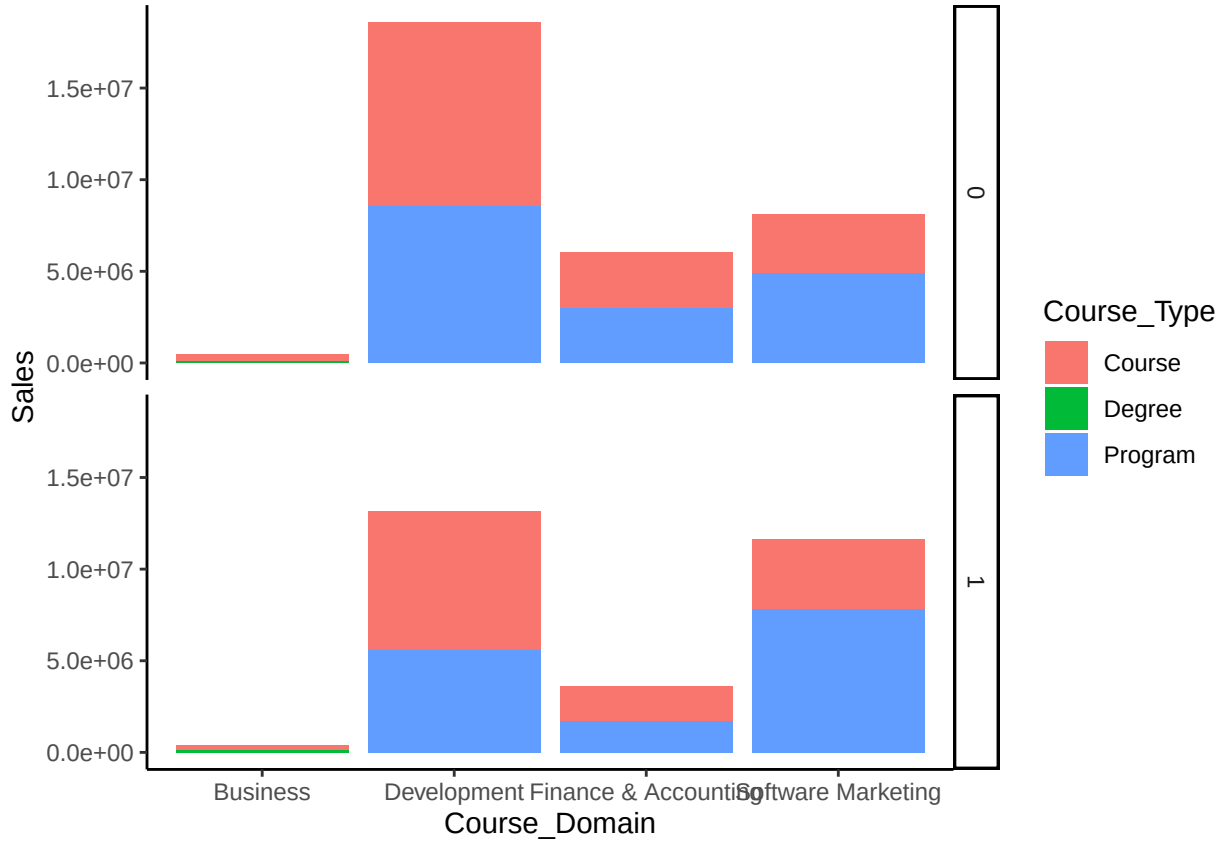
There are more sales for *Development Domain* courses and of *Course Type*, followed by *Software Marketting* for *Program Type*.

Analyze the Effect of short Term promotion,Long term Promotion and Public Holidays on sales

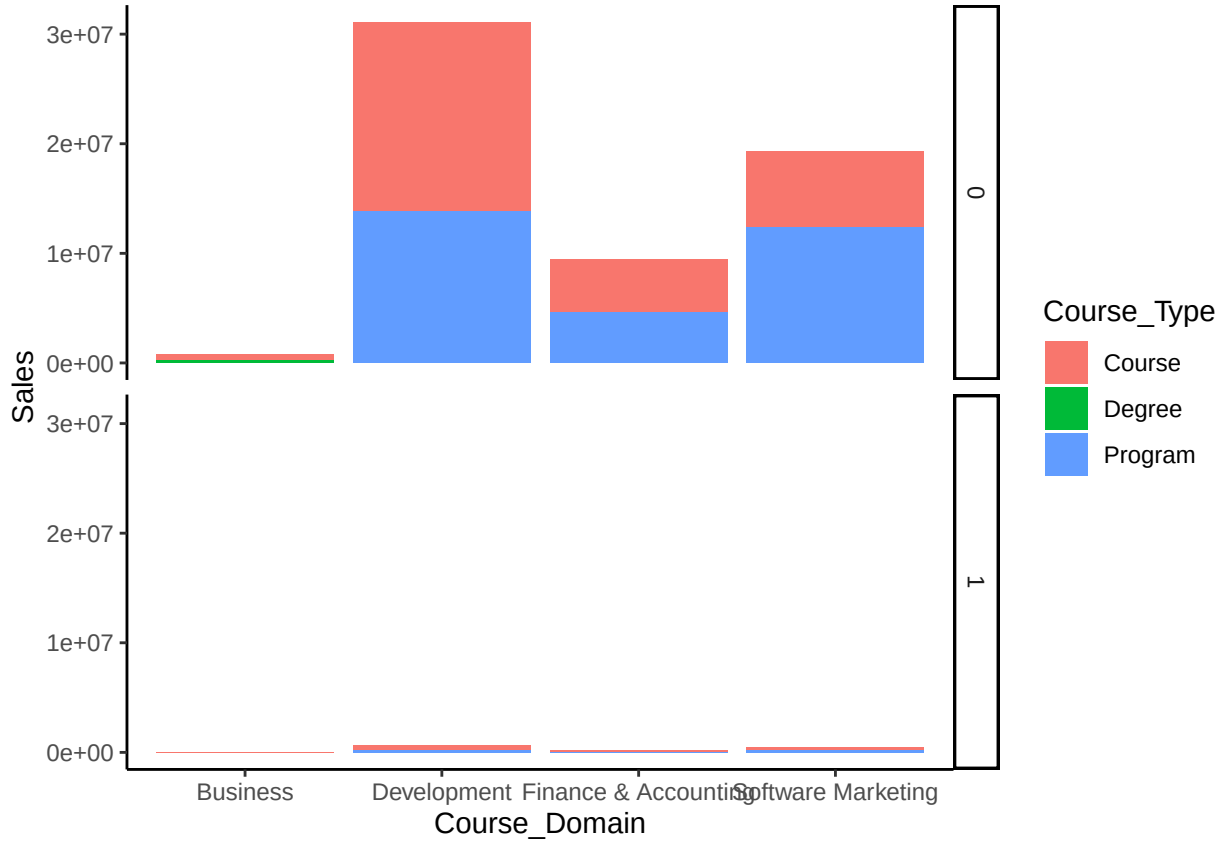
Short Promotion



Long Promotion



Public Holiday



From the above 3 graph we can see that Sale of courses is more when Short Promotion and Long Promotion is live and when there is Regional Holiday.

Model Building

For model Building we need to pre-process the data and after that Data splitting into traindata and test data.

Data Preprocessing

1- Imputation of missing values in the data.

2- Encoding of the Categorical features of the data.

```
## Classes 'data.table' and 'data.frame':   512087 obs. of  11 variables:
##  $ ID               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Day_No           : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Course_ID        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Course_Domain    : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ Course_Type      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Short_Promotion  : num  1 1 1 1 1 1 2 2 2 2 ...
##  $ Public_Holiday   : num  2 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ Long_Promotion    : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ User_Traffic      : int  11004 13650 11655 12054 6804 10395 16023 14385 16485 13377 ...
##  $ Competition_Metric: num  0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.007 0.007 ...
##  $ Sales             : int  81 79 75 80 41 62 122 114 121 100 ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "index")= int
##   ..- attr(*, "__Sales")= int  111857 112221 112225 135671 227991 228356 295634 295637 295640 318859
```

```
##        ID             Day_No        Course_ID     Course_Domain
##  Min.   :     1   Min.   :  1.0   Min.   :  1.0   Min.   :1.000
##  1st Qu.:136963   1st Qu.:214.0   1st Qu.:150.0   1st Qu.:2.000
##  Median :273984   Median :427.0   Median :300.0   Median :2.000
##  Mean   :274007   Mean   :434.9   Mean   :300.4   Mean   :2.791
##  3rd Qu.:411066   3rd Qu.:658.0   3rd Qu.:451.0   3rd Qu.:4.000
##  Max.   :548027   Max.   :882.0   Max.   :600.0   Max.   :4.000
##   Course_Type   Short_Promotion Public_Holiday  Long_Promotion
##  Min.   :1.00   Min.   :1.00    Min.   :1.000   Min.   :1.000
##  1st Qu.:1.00   1st Qu.:1.00    1st Qu.:1.000   1st Qu.:1.000
##  Median :1.00   Median :1.00    Median :1.000   Median :1.000
##  Mean   :1.97   Mean   :1.38    Mean   :1.032   Mean   :1.489
##  3rd Qu.:3.00   3rd Qu.:2.00    3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :3.00   Max.   :2.00    Max.   :2.000   Max.   :2.000
##   User_Traffic    Competition_Metric     Sales
##  Min.   :   168   Min.   :0.00000    Min.   :  0.0
##  1st Qu.: 10584   1st Qu.:0.01000    1st Qu.: 84.0
##  Median : 13776   Median :0.03500    Median :111.0
##  Mean   : 15375   Mean   :0.07335    Mean   :120.8
##  3rd Qu.: 18123   3rd Qu.:0.09400    3rd Qu.:146.0
##  Max.   :100002   Max.   :0.76800    Max.   :682.0
```

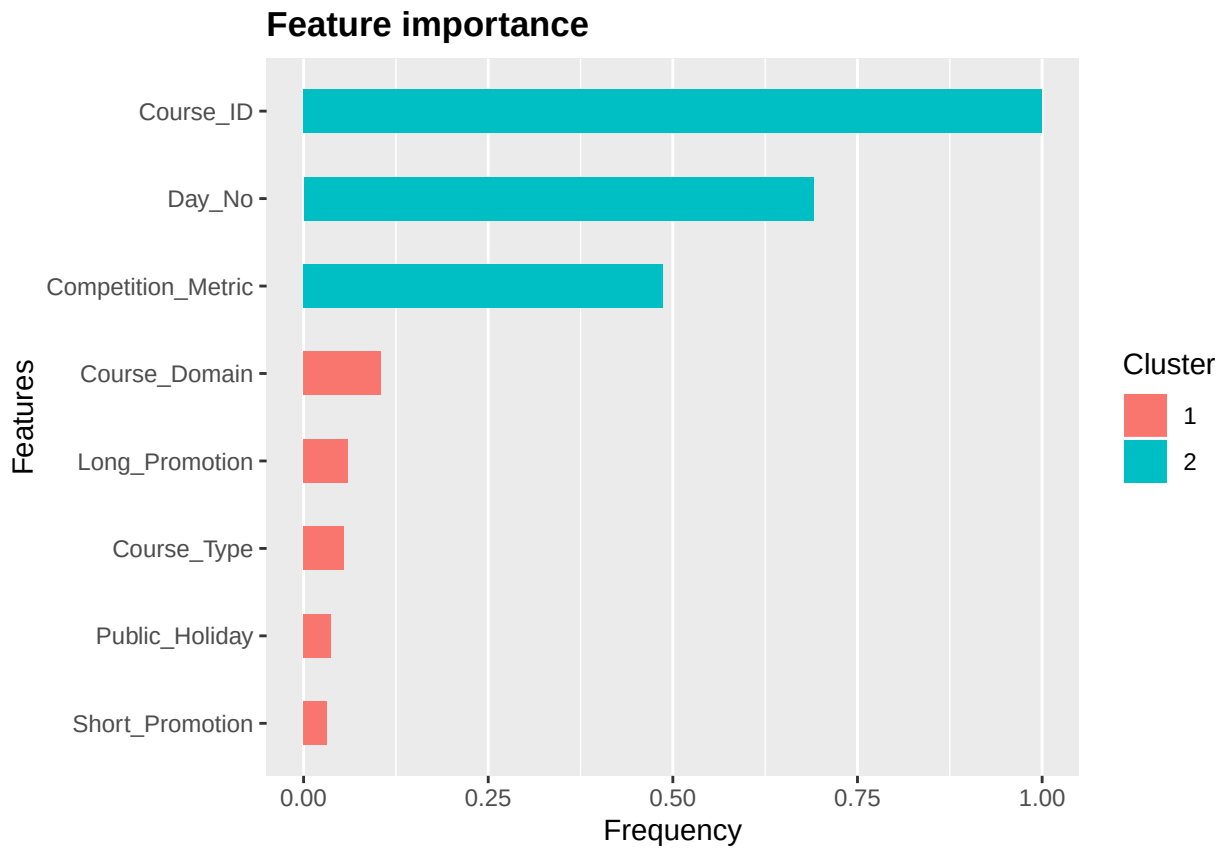
Predictive Model and Evaluation

Regression with Xgboost Model

Prediction on test data

```
##     ID Sales
## 1: 883   134
## 2: 884   129
## 3: 885   122
## 4: 886   122
## 5: 887    69
## 6: 888    69
```

Important Features

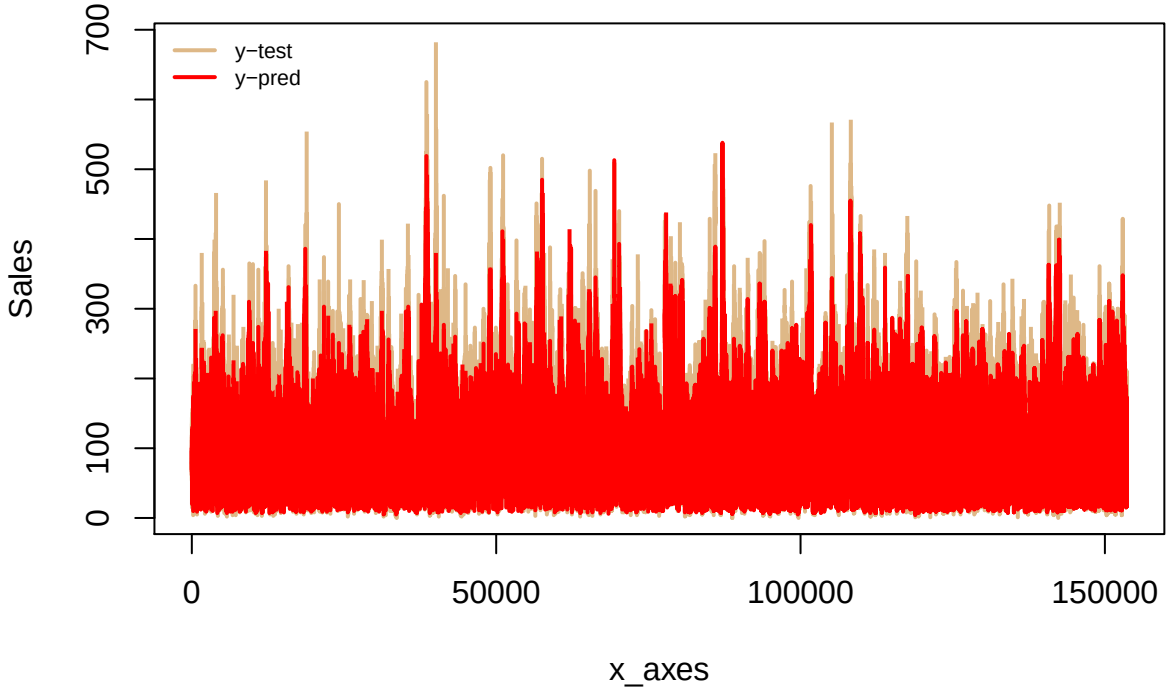


Evaluation

For evaluation of model we will use 1000*RMSLE where RMSLE is Root of Mean Squared Logarithmic Error.

```
## [1] 189.0225
```

Visualize Result



Regression with CNN Model

Extract the input dimension for the Keras model

```
## [1] 8 1
```

Model Fitting

```
## Model: "sequential"
## _________________________________________________________________________
## Layer (type)                    Output Shape                  Param #
## =========================================================================
## conv1d (Conv1D)                 (None, 7, 64)                 192
## _________________________________________________________________________
## conv1d_1 (Conv1D)               (None, 6, 64)                 8256
## _________________________________________________________________________
## conv1d_2 (Conv1D)               (None, 5, 64)                 8256
## _________________________________________________________________________
## conv1d_3 (Conv1D)               (None, 4, 64)                 8256
## _________________________________________________________________________
```

```
## flatten (Flatten)                   (None, 256)                   0
## _________________________________________________________________________
## dense (Dense)                       (None, 64)                    16448
## _________________________________________________________________________
## dense_1 (Dense)                     (None, 64)                    4160
## _________________________________________________________________________
## dense_2 (Dense)                     (None, 16)                    1040
## _________________________________________________________________________
## dense_3 (Dense)                     (None, 8)                     136
## _________________________________________________________________________
## dense_4 (Dense)                     (None, 1)                     9
## =========================================================================
## Total params: 46,753
## Trainable params: 46,753
## Non-trainable params: 0
## _________________________________________________________________________
```

```
##     loss
## 892.3967
```

Prediction on test data

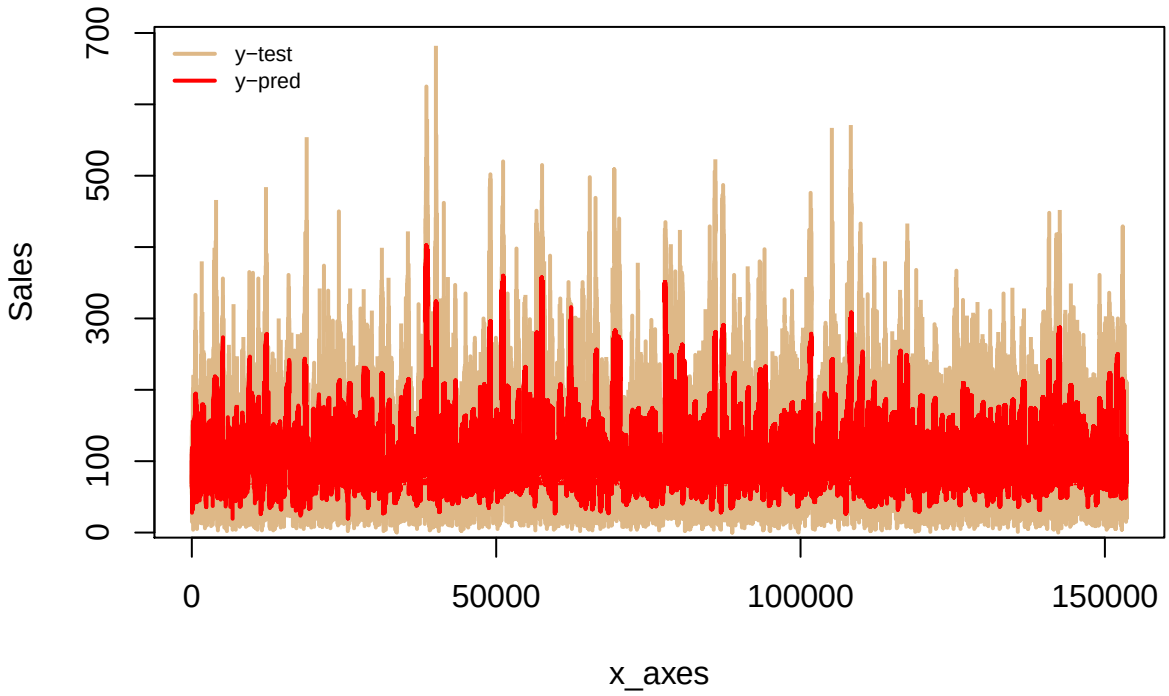
```
##     ID Sales
## 1: 883   115
## 2: 884   113
## 3: 885   125
## 4: 886   122
## 5: 887    61
## 6: 888    62
```

Evaluation Metric

To evaluate our model we will use 1000*RMSLE where RMSLE is Root of Mean Squared Logarithmic Error.

```
## [1] 278.7996
```

Visualize Result



Ensemble Model

Combine the result of all best models.

```
##      ID Sales
## 1: 883   124
## 2: 884   121
## 3: 885   120
## 4: 886   119
## 5: 887    66
## 6: 888    66
```