

BUSINESS REPORT

Capstone Project Notes 2

+

•

o

RAMESH PANDEY

PGP-DSBA Online

May' 23

Date: 16/07/2023

Table of Contents

List of Tables	4
Questions:	+
1. Model building and interpretation	
1.a. Build various models.....	6-11, 15-18, 22-25, 29-32
1.b. Test your predictive model.....	12-13, 19-20, 26-27, 33-34
1.c. Interpretation of the model.....	14, 21, 28, 35



Table of Contents

- 2. Model Tuning and business implication
 - 2.a. Model Tuning.....38 - 53
 - 2.b. Ensemble modelling.....54 - 68
 - 2.c. Any other model tuning.....69 - 76
 - 2.d. Interpretation of the most optimum model.....77 - 85



Tables 1: Coefficient or Importance of Features.....10, 11, 17, 18, 24,
25, 31, 32, 38, 39, 43, 44, 50, 51, 57, 58, 64, 65, 72, 73

Tables 2: Model Score/ Accuracy.....13, 20, 27, 34, 40, 45, 52, 59,
66, 67, 74, 75, 77, 78

Question 1 - Model building and interpretation



a) Build various models



I have built various models to predict the outcomes.

To build the models, first I have split the data into Train and Test set with the train size as 0.7 of the total data and Test data as 0.3 of total data.

Shape of the data are as:

Input Train - (3164, 51)

Input Test - (1356, 51)

Output Train - (3164,)

Output Test - (1356,)

a) Build various models

i. Linear Regression:



The first model built is Linear Regression model as the output is continuous data type.

Linear regression model was imported from sklearn.

The model was fit on train data and for the numerical column the coefficient was found to be highest for the SumAssured column and the value was 823.1767.

Followed by CustTensure having value of 206.6318.

a) Build various models

i. Linear Regression:



The next numerical column having higher coefficient is of Age having the value of 190.9338.

In the categorical column the highest coefficient is for Designation_VP is 425.4608.

Followed by Occupation_Free Lancer having value of 347.5439.

There are many features having negative coefficients which means that it has negative relation with the target column which is AgentBonus.

a) Build various models

i. Linear Regression:

The numerical column having negative coefficient are LastMonthCalls having value of -14.7817 and NumberOfPolicy having value of -5.4919.

Some of categorical column having negative values in decreasing order of magnitude are Designation_Executive having value of -287.4425, Designation_Manager having value of -276.31366, etc.

The values can be seen in the next slide.

- o

a) Build various models

I. Linear Regression:

```
The coefficient for Age is 190.9337899414458
The coefficient for CustTenure is 206.63181795669516
The coefficient for NumberOfPolicy is -5.491874259658459
The coefficient for MonthlyIncome is 128.18740542812228
The coefficient for ExistingPolicyTenure is 106.69351387830571
The coefficient for SumAssured is 823.1766552471993
The coefficient for LastMonthCalls is -14.781758659210254
The coefficient for Channel_Agent is -21.823961634864133
The coefficient for Channel_Online is 41.043214858879686
The coefficient for Channel_Third Party Partner is -19.21925322401632
The coefficient for Occupation_Free Lancer is 347.5439488456005
The coefficient for Occupation_Large Business is -60.76896186432796
The coefficient for Occupation_Salaried is -105.77432635353921
The coefficient for Occupation_Small Business is -181.00066062773803
The coefficient for EducationField_Diploma is 40.22369334501558
The coefficient for EducationField_Engineer is -71.36135899743928
The coefficient for EducationField_Graduate is -10.618113711873896
The coefficient for EducationField_MBA is 41.43146232110256
The coefficient for EducationField_Post Graduate is -56.86201057110724
The coefficient for EducationField_Under Graduate is 57.18632761430302
The coefficient for Gender_Female is -6.252512583612166
The coefficient for Gender_Male is 6.2525125836106135
The coefficient for ExistingProdType_1 is 14.077579964000677
The coefficient for ExistingProdType_2 is 27.915657451657864
The coefficient for ExistingProdType_3 is -117.43237660307292
The coefficient for ExistingProdType_4 is -28.1943754647028
The coefficient for ExistingProdType_5 is 65.63865923439093
The coefficient for ExistingProdType_6 is 37.99485541772503
```

a) Build various models

I. Linear Regression:

```
The coefficient for Designation_AVP is 241.24327452146653
The coefficient for Designation_Executive is -287.4425608169886
The coefficient for Designation_Manager is -276.3136672459488
The coefficient for Designation_Senior Manager is -102.94785577655932
The coefficient for Designation_VP is 425.46080931803544
The coefficient for MaritalStatus_Divorced is 20.339937916088175
The coefficient for MaritalStatus_Married is -29.153286482502768
The coefficient for MaritalStatus_Single is 8.813348566417272
The coefficient for Complaint_0 is -18.672525009462944
The coefficient for Complaint_1 is 18.672525009462973
The coefficient for Zone_East is -32.85965828677345
The coefficient for Zone_North is -60.57036129543545
The coefficient for Zone_South is 151.24348695926614
The coefficient for Zone_West is -57.81346737705249
The coefficient for PaymentMethod_Half Yearly is 2.910976206924687
The coefficient for PaymentMethod_Monthly is 2.5330206126742354
The coefficient for PaymentMethod_Quarterly is 96.91562086113693
The coefficient for PaymentMethod_Yearly is -102.35961768073611
The coefficient for CustCareScore_1.0 is -3.8895126329727816
The coefficient for CustCareScore_2.0 is -4.217368939920576
The coefficient for CustCareScore_3.0 is 0.9545309993784006
The coefficient for CustCareScore_4.0 is 13.065186420457081
The coefficient for CustCareScore_5.0 is -5.912835846940613
```

b) Test your predictive model

i. Linear Regression:

I have tested the model using predict option on the test data.

Some of the output which I have got are:

([4319.91915063, 3359.01129159, 5293.03361832,..., 3270.75690795, 3608.5746284 , 3667.35563204])

The error for Linear regression model calculated through Mean squared error is 407518.40285.

The R square or accuracy score of the model for train and test data set are 0.80208 and 0.7986 respectively.

b) Test your predictive model

i. Linear Regression:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986

o

c) Interpretation of the model

i. Linear Regression:

Looking at the score I can see that the model is not that bad. It has given a decent accuracy for both train and test set.

There seems to be no overfitting in the model. The accuracy can be further improved though.



a) Build various models

II. Decision Tree:



The second model built is Decision Tree Regressor.

Decision Tree regressor model was imported from sklearn. The model was fit on train data and for the numerical column the most important feature was found to be the SumAssured column and the value was 0.71899.

Followed by MonthlyIncome having value of 0.0665.

a) Build various models

II. Decision Tree:

The next numerical column having weightage is of Age having value of 0.06574 and then CustTenure having the value of 0.0624.

In the categorical column the highest weightage is for Designation_AVP is 0.00504.

Followed by CustCareScore_1.0 having value of 0.002803.

The least weightage is given to Occupation_Free Lancer having value of 8.317828267995846e^-08.

The values of the coefficient can be seen in the next slides.

a) Build various models

II. Decision Tree:

```
The coefficient for SumAssured is 0.7189933278067617
The coefficient for MonthlyIncome is 0.06653864142408254
The coefficient for Age is 0.06574484012222387
The coefficient for CustTenure is 0.062458703513239705
The coefficient for ExistingPolicyTenure is 0.011614424477351374
The coefficient for LastMonthCalls is 0.009605918450750707
The coefficient for NumberOfPolicy is 0.005956359890291309
The coefficient for Designation_AVP is 0.005041567223871081
The coefficient for CustCareScore_1.0 is 0.0028033793293497575
The coefficient for CustCareScore_3.0 is 0.0026077218047351586
The coefficient for Designation_VP is 0.0024065677732859145
The coefficient for MaritalStatus_Divorced is 0.002349618860012996
The coefficient for Zone_West is 0.002285412313198984
The coefficient for EducationField_Under Graduate is 0.0021666602384913477
The coefficient for Channel_Online is 0.0021332760920634336
The coefficient for ExistingProdType_3 is 0.00205640718312153
The coefficient for CustCareScore_5.0 is 0.0020350243132158134
The coefficient for ExistingProdType_4 is 0.001996379263800319
The coefficient for Designation_Executive is 0.0019873593664080967
The coefficient for MaritalStatus_Married is 0.0018281743104038334
The coefficient for Channel_Third Party Partner is 0.0018192867775996483
The coefficient for MaritalStatus_Single is 0.001685662442021655
The coefficient for EducationField_Engineer is 0.0014223355631627437
The coefficient for Gender_Female is 0.0013924043372140179
The coefficient for EducationField_Diploma is 0.001390405861697475
The coefficient for Designation_Senior Manager is 0.0013767760385090305
```

a) Build various models

II. Decision Tree:

```
The coefficient for CustCareScore_4.0 is 0.0013754053006796599
The coefficient for Complaint_1 is 0.0012643321961661165
The coefficient for EducationField_Graduate is 0.0012406511521741467
The coefficient for ExistingProdType_5 is 0.0011909736828704808
The coefficient for Designation_Manager is 0.0011589030434515644
The coefficient for EducationField_Post Graduate is 0.0011087676742638712
The coefficient for PaymentMethod_Half Yearly is 0.0010758826609598491
The coefficient for PaymentMethod_Yearly is 0.0009659788794421976
The coefficient for Complaint_0 is 0.0009371472646971479
The coefficient for Occupation_Salaried is 0.0008258077142017561
The coefficient for Occupation_Large Business is 0.0008009960744798074
The coefficient for Occupation_Small Business is 0.0006859223306360421
The coefficient for ExistingProdType_1 is 0.0006565546458746388
The coefficient for Channel_Agent is 0.0006293717649489699
The coefficient for Zone_North is 0.0006278443243897942
The coefficient for Zone_East is 0.000627062970232026
The coefficient for CustCareScore_2.0 is 0.0005895073464112644
The coefficient for ExistingProdType_2 is 0.0005616781826222215
The coefficient for EducationField_MBA is 0.0004989724464456227
The coefficient for ExistingProdType_6 is 0.0004900842793273084
The coefficient for Gender_Male is 0.00040274926485016376
The coefficient for PaymentMethod_Quarterly is 0.0002890551160344845
The coefficient for PaymentMethod_Monthly is 0.00015552648019725798
The coefficient for Zone_South is 0.00014410724949662075
The coefficient for Occupation_Free Lancer is 8.317828267995846e-08
```

b) Test your predictive model

II. Decision Tree:

I have tested the model using predict option on the test data.

+ .

Some of the output which I have got are:
([4449., 3874., 5995., ..., 4207., 3118., 2998.])

The error for Decision Tree model calculated through Mean squared error is 571543.962389.

The R square or accuracy score of the model for train and test data set are 1.0 and 0.717469 respectively.

o

b) Test your predictive model

II. Decision Tree:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175

o

c) Interpretation of the model

II. Decision Tree:

Looking at the score I can see that the model is overfitting as the⁺ Train score is 1 and the test score is 0.72 which shows that the model has performed well on the train data but has not performed that well on the test data.

The accuracy can be further improved.

Some tuning needs to be done in order to achieve better accuracy.

a) Build various models

III. Random Forest:



The third model built is Random Forest Regressor which is an ensemble model.

- Random Forest regressor model was imported from sklearn.
- The model was fit on train data and for the numerical column the most important feature was found to be the SumAssured column and the value was 0.72159.
- Followed by Age having value of 0.06958.

a) Build various models

III. Random Forest:

The next numerical column having weightage is of **MonthlyIncome** having value of 0.06507 and then **CustTenure** having the value of 0.0606.

In the categorical column the highest weightage is for **Designation_AVP** is 0.00332.

Followed by **ExistingProdType_3** having value of 0.002019.

The least weightage is given to **Occupation_Free Lancer** having value of 9.584370050111113e^-07.

The values of the coefficient can be seen in the next slides.

a) Build various models

III. Random Forest:

```
The coefficient for SumAssured is 0.721591907099885
The coefficient for Age is 0.0695849257818945
The coefficient for MonthlyIncome is 0.06507169089662587
The coefficient for CustTenure is 0.060601470605676215
The coefficient for ExistingPolicyTenure is 0.012121084541226729
The coefficient for LastMonthCalls is 0.010901569781192862
The coefficient for NumberOfPolicy is 0.005688114615067856
The coefficient for Designation_AVP is 0.003321313471244778
The coefficient for ExistingProdType_3 is 0.0020197087687634004
The coefficient for MaritalStatus_Divorced is 0.002012905117369017
The coefficient for CustCareScore_5.0 is 0.0020101006274154164
The coefficient for CustCareScore_1.0 is 0.0019018373031129035
The coefficient for MaritalStatus_Married is 0.001778480674778494
The coefficient for Channel_Online is 0.0017687073177638367
The coefficient for MaritalStatus_Single is 0.0017479771226582158
The coefficient for CustCareScore_3.0 is 0.0017279959591595573
The coefficient for ExistingProdType_4 is 0.0016932405897749792
The coefficient for CustCareScore_4.0 is 0.0016122571694887713
The coefficient for EducationField_Diploma is 0.0015885014705880286
The coefficient for EducationField_Under Graduate is 0.0015764548551331526
The coefficient for Designation_Senior Manager is 0.0015460224618965108
The coefficient for Designation_Manager is 0.0014866703002487977
The coefficient for Designation_VP is 0.0014325900665232711
The coefficient for EducationField_Graduate is 0.0014124965670788718
The coefficient for Channel_Third Party Partner is 0.0013767132024570124
The coefficient for CustCareScore_2.0 is 0.0013475987825033661
```

a) Build various models

III. Random Forest:

```
The coefficient for ExistingProdType_5 is 0.0013376437563603588
The coefficient for Zone_North is 0.001336045733885983
The coefficient for PaymentMethod_Half Yearly is 0.0013232011498124207
The coefficient for PaymentMethod_Yearly is 0.0013093530443183188
The coefficient for Occupation_Salaried is 0.0012866784213833135
The coefficient for Channel_Agent is 0.0012632402106139149
The coefficient for Zone_West is 0.0012213932360887452
The coefficient for Gender_Female is 0.0011849839382152964
The coefficient for Occupation_Small Business is 0.0011575065584681994
The coefficient for Complaint_0 is 0.0011179107867336157
The coefficient for Complaint_1 is 0.0010880608124445605
The coefficient for Designation_Executive is 0.0010675751856712452
The coefficient for Gender_Male is 0.0010494302767880199
The coefficient for EducationField_Post Graduate is 0.0008527605244586597
The coefficient for EducationField_Engineer is 0.0007370631318625294
The coefficient for ExistingProdType_2 is 0.0006907963013588809
The coefficient for Occupation_Large Business is 0.0006504032727846318
The coefficient for PaymentMethod_Monthly is 0.0006087019995388354
The coefficient for Zone_East is 0.000504562840724458
The coefficient for ExistingProdType_1 is 0.00038628375738832677
The coefficient for EducationField_MBA is 0.0003423663119934767
The coefficient for ExistingProdType_6 is 0.0003211782100783365
The coefficient for PaymentMethod_Quarterly is 0.00019399642228705566
The coefficient for Zone_South is 4.5630919278914745e-05
The coefficient for Occupation_Free Lancer is 9.584370050111113e-07
```

b) Test your predictive model

III. Random Forest:

I have tested the model using predict option on the test data.

+ .

Some of the output which I have got are:

([4310.25, 3606.37, 5031.15, ..., 3506.55, 3799.02, 3405.6])

The error for Random Forest model calculated through Mean squared error is 290651.9734168879.

The R square or accuracy score of the model for train and test data set are 0.97875 and 0.8563 respectively.

o

b) Test your predictive model

III. Random Forest:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161
10	Random_Forest1	model_rf	0.9788	0.8563	290651.9734	0.8563

c) Interpretation of the model

III. Random Forest:

Looking at the score I can see that the model is overfitting as the⁺ Train score is 0.9788 and the test score is 0.8563 which shows that the model has performed well on the train data but has not performed that well on the test data.

The accuracy is decent and can be further improved.

Some tuning needs to be done in order to achieve better accuracy.

a) Build various models

IV. Gradient Boost:



The fourth model built is Gradient Boost Regressor which is an ensemble model.

Gradient Boost regressor model was imported from sklearn. The model was fit on train data and for the numerical column the most important feature was found to be the SumAssured column and the value was 0.76958. Followed by MonthlyIncome having value of 0.07897.

a) Build various models

IV. Gradient Boost:

The next numerical column having weightage is of Age having value of 0.06848 and then CustTenure having the value of 0.06122.

In the categorical column the highest weightage is for Designation_AVP is 0.003328.

Followed by Designation_Senior Manager having value of 0.000703.

There are many features having value of 0 like Occupation_Free Lancer, Zone_East, PaymentMethod_Yearly, etc.

The values of the coefficient can be seen in the next slides.

a) Build various models

IV. Gradient Boost:

```
The coefficient for SumAssured is 0.7695845016988022
The coefficient for MonthlyIncome is 0.07897699216913717
The coefficient for Age is 0.06848074974364164
The coefficient for CustTenure is 0.06122006959671558
The coefficient for ExistingPolicyTenure is 0.009328063687949343
The coefficient for Designation_AVP is 0.003328613112828938
The coefficient for LastMonthCalls is 0.0024154355393595363
The coefficient for Designation_Senior Manager is 0.0007034355214911852
The coefficient for Designation_Manager is 0.0005803653740292681
The coefficient for ExistingProdType_3 is 0.0005206342813539473
The coefficient for MaritalStatus_Divorced is 0.00041424688823554474
The coefficient for CustCareScore_4.0 is 0.0003599708410140271
The coefficient for Designation_VP is 0.0003496227129888212
The coefficient for EducationField_Diploma is 0.0003300413324426208
The coefficient for Occupation_Salaried is 0.0002911535381931539
The coefficient for NumberOfPolicy is 0.000290888878822656
The coefficient for CustCareScore_2.0 is 0.0002717178544693154
The coefficient for Channel_Agent is 0.0002669509163738145
The coefficient for Channel_Online is 0.00025031084707167475
The coefficient for Gender_Female is 0.0002321002453186777
The coefficient for Channel_Third Party Partner is 0.00023001704231893186
The coefficient for CustCareScore_3.0 is 0.00019049099057051597
The coefficient for PaymentMethod_Quarterly is 0.00018929897720341117
The coefficient for EducationField_MBA is 0.0001480901857348373
The coefficient for MaritalStatus_Single is 0.00013974380536781392
The coefficient for Zone_West is 0.00012473838263957055
```

a) Build various models

IV. Gradient Boost:

```
The coefficient for CustCareScore_5.0 is 9.735059666699295e-05
The coefficient for ExistingProdType_1 is 9.196308886940276e-05
The coefficient for EducationField_Under Graduate is 8.051100629915335e-05
The coefficient for PaymentMethod_Monthly is 7.546086265537179e-05
The coefficient for EducationField_Graduate is 7.288921251751608e-05
The coefficient for Occupation_Small Business is 6.939094819082337e-05
The coefficient for Designation_Executive is 6.632166906141917e-05
The coefficient for ExistingProdType_2 is 5.4895001624287106e-05
The coefficient for CustCareScore_1.0 is 5.134447318877103e-05
The coefficient for Zone_North is 4.0593659584743725e-05
The coefficient for MaritalStatus_Married is 3.256811959988008e-05
The coefficient for Complaint_0 is 3.057403412368905e-05
The coefficient for Complaint_1 is 6.6020463334482396e-06
The coefficient for PaymentMethod_Half Yearly is 6.240953194892007e-06
The coefficient for Gender_Male is 5.050164015277115e-06
The coefficient for Occupation_Free Lancer is 0.0
The coefficient for Occupation_Large Business is 0.0
The coefficient for EducationField_Engineer is 0.0
The coefficient for EducationField_Post Graduate is 0.0
The coefficient for ExistingProdType_4 is 0.0
The coefficient for ExistingProdType_5 is 0.0
The coefficient for ExistingProdType_6 is 0.0
The coefficient for Zone_East is 0.0
The coefficient for Zone_South is 0.0
The coefficient for PaymentMethod_Yearly is 0.0
```

b) Test your predictive model

IV. Gradient Boost:

I have tested the model using predict option on the test data.

+
•

Some of the output which I have got are:

([4567.83798589, 3302.959075 , 5129.22005981, ..., 3468.16167765, 3648.49245638, 3474.36918772])

The error for Gradient Boost model calculated through Mean squared error is 305311.503678.

The R square or accuracy score of the model for train and test data set are 0.8727 and 0.8491 respectively.

b) Test your predictive model

IV. Gradient Boost:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161
10	Random_Forest1	model_rf	0.9788	0.8563	290651.9734	0.8563
11	Random_Forest2	model_rf1	0.9790	0.8564	290534.3749	0.8564
12	Random_Forest3	model_rf2	0.8374	0.8213	361430.6817	0.8213
13	Random_Forest4	model_rf3	0.8375	0.8213	361478.2054	0.8213
14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350
15	Random_Forest6	model_rf5	0.8588	0.8350	333686.0254	0.8350
16	Gradient_Boost1	model_gb	0.8727	0.8491	305311.5037	0.8491

c) Interpretation of the model

IV. Gradient Boost:

Looking at the score I can see that there is no overfitting as the⁺ Train score is 0.8727 and the test score is 0.8491 which shows that the model has performed well on both the train data as well on the test data.

The accuracy is decent and can be tried to further improve.

Some tuning needs to be done in order to achieve better accuracy.

Question 2 - Model Tuning and business implication



a) Model Tuning

Ridge Regression:



Linear Regression model is tuned through Ridge model. So, I have used Ridge first.

The coefficient seem to be similar to the linear regression and their is not much changes in the score or the error.

MSE = 407507.8720320885

R2 = 0.7986 which is approx same as that of Linear regression.

So, their is not much of a change as compared to linear regression.

a) Model Tuning

Ridge Regression:

```
The coefficient for SumAssured is 823.1130251037838
The coefficient for Designation_VP is 423.9175477969683
The coefficient for Occupation_Free Lancer is 311.3406880754874
The coefficient for Designation_Executive is -286.34833740037885
The coefficient for Designation_Manager is -275.66952631380025
The coefficient for Designation_AVP is 240.83036035155976
The coefficient for CustTenure is 206.65488261200846
The coefficient for Age is 191.0017082708138
The coefficient for Occupation_Small Business is -167.7793685588469
The coefficient for Zone_South is 143.0142542018815
The coefficient for MonthlyIncome is 128.73427534525993
The coefficient for ExistingProdType_3 is -117.03995926213497
The coefficient for ExistingPolicyTenure is 106.68375678071926
The coefficient for Designation_Senior Manager is -102.73004444325368
The coefficient for PaymentMethod_Yearly is -101.6068078241971
The coefficient for PaymentMethod_Quarterly is 96.21825953783828
The coefficient for Occupation_Salaried is -94.3242182524084
The coefficient for EducationField_Engineer is -70.90930016046482
The coefficient for ExistingProdType_5 is 65.15320206104766
The coefficient for Zone_North is -57.83866103917779
The coefficient for EducationField_Under Graduate is 56.01552780093921
The coefficient for EducationField_Post Graduate is -55.737428066304396
The coefficient for Zone_West is -55.15868019507484
The coefficient for Occupation_Large Business is -49.23710126620409
The coefficient for EducationField_MBA is 41.741695015407664
The coefficient for Channel_Online is 41.01127256099513
The coefficient for EducationField_Diploma is 38.97789828438936
```

a) Model Tuning

Ridge Regression:

```
The coefficient for ExistingProdType_6 is 37.204049602011935
The coefficient for Zone_East is -30.016912965490043
The coefficient for MaritalStatus_Married is -29.174694682277924
The coefficient for ExistingProdType_4 is -28.287620532604368
The coefficient for ExistingProdType_2 is 28.265807431028403
The coefficient for Channel_Agent is -21.765539571909333
The coefficient for MaritalStatus_Divorced is 20.323020523733703
The coefficient for Channel_Third Party Partner is -19.245732989138517
The coefficient for Complaint_0 is -18.698947893978062
The coefficient for Complaint_1 is 18.698947893963865
The coefficient for LastMonthCalls is -14.73179691156285
The coefficient for ExistingProdType_1 is 14.704520699719062
The coefficient for CustCareScore_4.0 is 13.025273731096936
The coefficient for EducationField_Graduate is -10.08839287361226
The coefficient for MaritalStatus_Single is 8.851674158764448
The coefficient for Gender_Male is 6.257583638648606
The coefficient for Gender_Female is -6.257583638474835
The coefficient for CustCareScore_5.0 is -5.928531016161623
The coefficient for NumberOfPolicy is -5.479540375652404
The coefficient for CustCareScore_2.0 is -4.235427664102981
The coefficient for CustCareScore_1.0 is -3.928871117420558
The coefficient for PaymentMethod_Half Yearly is 3.0711319918061624
The coefficient for PaymentMethod_Monthly is 2.3174162943194987
The coefficient for CustCareScore_3.0 is 1.0675560667962696
```

c) Interpretation of the model

Ridge Regression:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986

o

c) Interpretation of the model

Ridge Regression:

Looking at the score I can see that there is no overfitting as the⁺ Train score is 0.8021 and the test score is 0.7986 which shows that the model has performed well on both the train data as well on the test data.

The accuracy is similar to linear regression model.



a) Model Tuning

Lasso Regression:



Linear Regression model is tuned through another model, Lasso model. So,

The coefficient seem to be similar to the linear regression and their is not much changes in the score or the error.

MSE = 407256.618276685

R2 = 0.79868 which is approx same as that of Linear regression.

So, their is not much of a change as compared to linear regression.

a) Model Tuning

Lasso Regression:

```
The coefficient for SumAssured is 823.235501676041
The coefficient for Designation_VP is 522.0930629144451
The coefficient for Designation_AVP is 341.3156617690332
The coefficient for Occupation_Free Lancer is 291.0623535393876
The coefficient for CustTenure is 206.74950742852528
The coefficient for Age is 191.10950811838217
The coefficient for Designation_Executive is -180.364978014672
The coefficient for Designation_Manager is -170.91698174287546
The coefficient for ExistingProdType_3 is -141.722615892842
The coefficient for MonthlyIncome is 130.58498690857041
The coefficient for Zone_South is 128.627236757605
The coefficient for ExistingPolicyTenure is 106.59656138100199
The coefficient for PaymentMethod_Yearly is -100.9098370433428
The coefficient for PaymentMethod_Quarterly is 87.87096876228068
The coefficient for EducationField_Post Graduate is -65.52001203532808
The coefficient for Channel_Online is 59.71334095951565
The coefficient for Occupation_Small Business is -57.49778355274091
The coefficient for ExistingProdType_4 is -55.415490622515364
The coefficient for EducationField_Engineer is -47.07018195171997
The coefficient for MaritalStatus_Married is -37.81215678006672
The coefficient for Complaint_0 is -37.033984869509176
The coefficient for ExistingProdType_5 is 34.692905982107234
The coefficient for EducationField_Under Graduate is 27.71359922828238
The coefficient for EducationField_MBA is 22.70809766575744
The coefficient for EducationField_Graduate is -22.525903135223093
The coefficient for Zone_East is 17.384680310612715
The coefficient for CustCareScore_4.0 is 16.05513161266601
```

a) Model Tuning

Lasso Regression:

```
The coefficient for LastMonthCalls is -14.376043414197122
The coefficient for Gender_Female is -12.104950422025025
The coefficient for MaritalStatus_Divorced is 10.949536313511729
The coefficient for EducationField_Diploma is 9.472126764104008
The coefficient for ExistingProdType_1 is -9.374410446928168
The coefficient for Occupation_Large Business is 8.191307368845978
The coefficient for NumberOfPolicy is -5.161566600049178
The coefficient for CustCareScore_3.0 is 4.968056468025464
The coefficient for ExistingProdType_6 is 3.1054687131443055
The coefficient for Zone_North is -3.0221142956539837
The coefficient for Channel_Agent is -2.128127414227892
The coefficient for CustCareScore_5.0 is -1.5835381517730618
The coefficient for Zone_West is -0.3348514533352945
The coefficient for Channel_Third Party Partner is 0.0
The coefficient for Occupation_Salaried is -0.0
The coefficient for Gender_Male is 0.0
The coefficient for ExistingProdType_2 is 0.0
The coefficient for Designation_Senior Manager is -0.0
The coefficient for MaritalStatus_Single is -0.0
The coefficient for Complaint_1 is 0.0
The coefficient for PaymentMethod_Half Yearly is 0.0
The coefficient for PaymentMethod_Monthly is -0.0
The coefficient for CustCareScore_1.0 is -0.0
The coefficient for CustCareScore_2.0 is -0.0
```

c) Interpretation of the model

Lasso Regression:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987

o

c) Interpretation of the model

Lasso Regression:

Looking at the score I can see that there is no overfitting as the Train score is 0.8021 and the test score is 0.7987 which shows that the model has performed well on both the train data as well on the test data.

The accuracy is similar to linear regression model.

Both Ridge and Lasso have similar error and accuracy for train and test data as compared to Linear regression model and there is no much improvement of the data.

a) Model Tuning

Decision Tree:

There are different parameters to tune the decision tree model such as changing max_depth and minimum sample split.

First I have chosen max_depth as 3.

The model train and test score are 0.7342 and 0.7195 respectively.

This model reduces the overfitting in the initial model but doesn't increase the accuracy score of the model. Hence, I have tried other max_depth as 5.

The model train and test score are 0.8132 and respectively which gives a better accuracy for both train and test data.

a) Model Tuning

Decision Tree:

The accuracy of the model increases for both train and test data by increasing the max_depth upto certain point.⁺ After max_depth of 8 overfitting increases and also accuracy decreases.

The optimum max_depth value came out to be 7 for which the accuracy score is 0.8568 and 0.8161 for train and test data respectively.

Changing the min sample split value doesn't seem to have effect on the accuracy of the model.

Hence, the fine tuned model of Decision Tree has max_depth of 7 and other parameters remains to be their default values.

a) Model Tuning

Decision Tree:

The model name for the best Decision Tree model is `model_dt_6` having the accuracy of 0.8568 and 0.8161 and error of 371980.1317.

The most important feature of the model is `SumAssured` followed by `MonthlyIncome` then `Age`.

`Designation_AVP` is the most important categorical feature for the model followed by `ExistingProdType_3`.

There is no overfitting or underfitting in the data.

There are many features whose value is 0, that is they are not so important.

The value for the coefficients can be seen in the next few slides.

a) Model Tuning

Decision Tree:

```
The coefficient for SumAssured is 0.8174494811471346
The coefficient for MonthlyIncome is 0.06392887850417361
The coefficient for Age is 0.0523796043315608
The coefficient for CustTenure is 0.048034030350285147
The coefficient for Designation_AVP is 0.0031047612414513434
The coefficient for ExistingPolicyTenure is 0.0028984573053714327
The coefficient for Designation_VP is 0.002267500100956671
The coefficient for LastMonthCalls is 0.001837586257078765
The coefficient for ExistingProdType_3 is 0.0015609078072864571
The coefficient for Designation_Executive is 0.0012153304895738684
The coefficient for EducationField_Under Graduate is 0.0008581134596461498
The coefficient for CustCareScore_3.0 is 0.0008342208515567158
The coefficient for MaritalStatus_Single is 0.0006366873677834484
The coefficient for ExistingProdType_4 is 0.0005294759031389812
The coefficient for EducationField_MBA is 0.00047491321149415274
The coefficient for Occupation_Small Business is 0.00033784707840037105
The coefficient for Designation_Manager is 0.0003333164961246227
The coefficient for PaymentMethod_Yearly is 0.000318750822877247
The coefficient for MaritalStatus_Divorced is 0.00021989246148660955
The coefficient for NumberOfPolicy is 0.0001899930055837531
The coefficient for Zone_East is 0.00018925120334110668
The coefficient for Channel_Online is 0.00013390200381930824
The coefficient for MaritalStatus_Married is 8.901519901029109e-05
The coefficient for Gender_Female is 8.718234459627025e-05
The coefficient for CustCareScore_1.0 is 7.68930530112761e-05
The coefficient for EducationField_Diploma is 1.175544906283262e-05
The coefficient for PaymentMethod_Half Yearly is 1.965782743559759e-06
```

a) Model Tuning

Decision Tree:

```
The coefficient for Zone_West is 2.8677145058199606e-07
The coefficient for Channel_Agent is 0.0
The coefficient for Channel_Third Party Partner is 0.0
The coefficient for Occupation_Free Lancer is 0.0
The coefficient for Occupation_Large Business is 0.0
The coefficient for Occupation_Salaried is 0.0
The coefficient for EducationField_Engineer is 0.0
The coefficient for EducationField_Graduate is 0.0
The coefficient for EducationField_Post Graduate is 0.0
The coefficient for Gender_Male is 0.0
The coefficient for ExistingProdType_1 is 0.0
The coefficient for ExistingProdType_2 is 0.0
The coefficient for ExistingProdType_5 is 0.0
The coefficient for ExistingProdType_6 is 0.0
The coefficient for Designation_Senior Manager is 0.0
The coefficient for Complaint_0 is 0.0
The coefficient for Complaint_1 is 0.0
The coefficient for Zone_North is 0.0
The coefficient for Zone_South is 0.0
The coefficient for PaymentMethod_Monthly is 0.0
The coefficient for PaymentMethod_Quarterly is 0.0
The coefficient for CustCareScore_2.0 is 0.0
The coefficient for CustCareScore_4.0 is 0.0
The coefficient for CustCareScore_5.0 is 0.0
```

c) Interpretation of the model

Decision Tree:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161

c) Interpretation of the model

Decision Tree:

The best tuned Decision tree model is the model_dt_6 model⁺ with max_depth = 7.

The error for the model is the least among all the models till now and has given the highest accuracy for both train and test data and is higher than the Linear regression model or its tuned model of Ridge and Lasso model.

The error value is 371980.1317 and the train and test accuracy are 0.8568 and 0.8161 respectively.

- o

a) Model Tuning

Random Forest:

There are different parameters to tune the random forest model such as changing number of trees, max_depth and minimum sample split.

First I have chosen number of trees as 110. It has no effect from the basic Random Forest model where number of trees is 100.

Train and test accuracy are 0.9790 and 0.8564 where as the error value is 290534.3749.

There is clearly overfitting in the data.



a) Model Tuning

Random Forest:

For next tuning I have kept the number of observations as default which is 100 and increased `max_depth` to 5⁺.

I can see that the model score decreases a little and also overfitting in the train data decreases.

The error increases for this model to 361430.6817 and the train and test scores are 0.8374 and 0.8213 respectively.

Increasing the `max_depth` value increases the accuracy of the model and error decreases.

Error value is 333725.5135 and train and test scores are 0.8590 and 0.8350 respectively.

a) Model Tuning

Random Forest:

For the next model tuning I have kept n_estimator to be 100 and max_depth = 6 and changed the min_sample_split⁺ to 5.

But there is no much effect on the accuracy of the model, just the error has reduced a little.

Therefore, the best Random Forest model is model_rf5 and is the best model so far having the highest accuracy for train and test data as 0.8588 and 0.8350 respectively and also the least error of 333686.0254.

There is no overfitting or underfitting in the data.

The value for the coefficients can be seen in the next few slides.

a) Model Tuning

Random Forest:

```
The coefficient for SumAssured is 0.8236979450682067
The coefficient for Age is 0.05815328005216963
The coefficient for MonthlyIncome is 0.05698396702294351
The coefficient for CustTenure is 0.047729198115612485
The coefficient for ExistingPolicyTenure is 0.0022038755392990764
The coefficient for Designation_AVP is 0.002064170749543227
The coefficient for LastMonthCalls is 0.001412505900456007
The coefficient for Designation_VP is 0.0008434679717927004
The coefficient for ExistingProdType_3 is 0.0007542057119640893
The coefficient for Designation_Senior Manager is 0.0004631542979119003
The coefficient for Designation_Manager is 0.0004108696910678102
The coefficient for MaritalStatus_Single is 0.00040639412762329645
The coefficient for Designation_Executive is 0.00039853959983444926
The coefficient for Channel_Third Party Partner is 0.00035367004942977597
The coefficient for NumberOfPolicy is 0.0003420516179149526
The coefficient for Channel_Agent is 0.00023389588121656478
The coefficient for CustCareScore_1.0 is 0.00021050890434751401
The coefficient for MaritalStatus_Divorced is 0.00021045913920990781
The coefficient for CustCareScore_2.0 is 0.0001881917252243423
The coefficient for PaymentMethod_Half Yearly is 0.00018104223421846121
The coefficient for EducationField_Under Graduate is 0.00017126892637489506
The coefficient for CustCareScore_5.0 is 0.00016526928125906867
The coefficient for EducationField_Diploma is 0.00016154609049124795
The coefficient for MaritalStatus_Married is 0.00014598215936667802
The coefficient for PaymentMethod_Yearly is 0.00014477272098882443
The coefficient for Zone_West is 0.00014420225767186872
The coefficient for Zone_East is 0.00012302391531428784
```

a) Model Tuning

Random Forest:

```
The coefficient for CustCareScore_3.0 is 0.0001167805946347173
The coefficient for Gender_Female is 0.00011640741515983503
The coefficient for EducationField_Graduate is 0.0001117398813702444
The coefficient for ExistingProdType_5 is 0.00011010858828722682
The coefficient for Occupation_Salaried is 0.00010903542049837898
The coefficient for Gender_Male is 0.0001082792998225016
The coefficient for Complaint_1 is 0.0001047680977832921
The coefficient for Occupation_Small Business is 9.670927023898692e-05
The coefficient for EducationField_Post Graduate is 9.575481447009517e-05
The coefficient for CustCareScore_4.0 is 9.056049369368568e-05
The coefficient for ExistingProdType_2 is 9.043382190313311e-05
The coefficient for Zone_North is 8.420166179304643e-05
The coefficient for Occupation_Large Business is 8.184460316188578e-05
The coefficient for ExistingProdType_6 is 6.799296941563104e-05
The coefficient for Channel_Online is 6.598285891116774e-05
The coefficient for EducationField_Engineer is 5.5497608125437106e-05
The coefficient for EducationField_MBA is 5.492956927906034e-05
The coefficient for ExistingProdType_4 is 5.3972203376173145e-05
The coefficient for Complaint_0 is 3.478188358383777e-05
The coefficient for PaymentMethod_Monthly is 3.115731139009326e-05
The coefficient for ExistingProdType_1 is 1.2765328252764872e-05
The coefficient for PaymentMethod_Quarterly is 8.837553395614791e-06
The coefficient for Occupation_Free Lancer is 0.0
The coefficient for Zone_South is 0.0
```

c) Interpretation of the model

Random Forest:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161
10	Random_Forest1	model_rf	0.9788	0.8563	290651.9734	0.8563
11	Random_Forest2	model_rf1	0.9790	0.8564	290534.3749	0.8564
12	Random_Forest3	model_rf2	0.8374	0.8213	361430.6817	0.8213
13	Random_Forest4	model_rf3	0.8375	0.8213	361478.2054	0.8213
14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350
15	Random_Forest6	model_rf5	0.8588	0.8350	333686.0254	0.8350

c) Interpretation of the model

Random Forest:

The best tuned Decision tree model is the model_rf5 model with $n_{estimators}=100$, $max_depth = 6$ and $min_sample_split = 5$.

The error for the model is the least among all the models till now and has given the highest accuracy for both train and test data and is higher than the Linear regression model or its tuned model of Ridge and Lasso model and also Decision Tree.

The error value is 333686.0254 and the train and test accuracy are 0.8588 and 0.8350 respectively.

a) Model Tuning

Gradient Boost:

There are different parameters to tune the Gradient Boost model such as learning rate, changing number of trees, etc.

First I have chosen learning rate as 0.05. It has shown that there is not much change in the accuracy when the learning rate is 0.1, the train and test accuracy has decreased a little.

Train and test accuracy are 0.8566 and 0.8402 where as the error value is 323305.7346.

There is clearly no overfitting or underfitting in the data.

a) Model Tuning

Gradient Boost:

For next tuning I have changed the number of observations as 120 and keeping the learning rate as 0.05.

I can see that the model score increases a little. The error decreases for this model to 317730.7306 and the train and test scores are 0.8606 and 0.8429 respectively.



a) Model Tuning

Gradient Boost:

Therefore, the best tuned Gradient Boost model is model_gb2 and is the best model after model_gb model having the second highest accuracy for train and test data as 0.8606 and 0.8429 respectively and also the least error of 317730.7306.

There is no overfitting or underfitting in the data. The value for the coefficients can be seen in the next few slides.

- o

a) Model Tuning

Gradient Boost:

```
The coefficient for SumAssured is 0.7794326508455697
The coefficient for MonthlyIncome is 0.07539373822234338
The coefficient for Age is 0.06581134537904307
The coefficient for CustTenure is 0.06149548902723668
The coefficient for ExistingPolicyTenure is 0.00882681188729652
The coefficient for Designation_AVP is 0.0031904805071098804
The coefficient for LastMonthCalls is 0.0012902677822969415
The coefficient for Designation_Manager is 0.0006890487912144726
The coefficient for Designation_VP is 0.0006859921219588329
The coefficient for ExistingProdType_3 is 0.0005453365867529014
The coefficient for Designation_Senior Manager is 0.0004817780093187311
The coefficient for CustCareScore_4.0 is 0.0002457354516234591
The coefficient for EducationField_Graduate is 0.00020010212122227664
The coefficient for NumberOfPolicy is 0.00019422588761827694
The coefficient for CustCareScore_2.0 is 0.00018951404281870583
The coefficient for ExistingProdType_1 is 0.00014735318805798715
The coefficient for Channel_Third Party Partner is 0.00013433864792201336
The coefficient for MaritalStatus_Divorced is 0.00013080836936437383
The coefficient for Channel_Agent is 0.0001255282039935162
The coefficient for Designation_Executive is 0.00012053663417128082
The coefficient for Channel_Online is 0.00011987225983974544
The coefficient for Zone_West is 8.755817673683134e-05
The coefficient for ExistingProdType_2 is 7.308002667395532e-05
The coefficient for EducationField_Under Graduate is 7.197214357188711e-05
The coefficient for Occupation_Small Business is 7.095779999998843e-05
The coefficient for CustCareScore_3.0 is 6.205582991044318e-05
The coefficient for CustCareScore_1.0 is 5.772911197258921e-05
```

a) Model Tuning

Gradient Boost:

```
The coefficient for EducationField_Post Graduate is 4.084911717057931e-05
The coefficient for Occupation_Salaried is 3.144969577664363e-05
The coefficient for EducationField_MBA is 2.9625578658585938e-05
The coefficient for Zone_North is 1.8281476162260652e-05
The coefficient for PaymentMethod_Half Yearly is 3.4523620469304995e-06
The coefficient for Gender_Female is 2.0347145464781592e-06
The coefficient for Occupation_Free Lancer is 0.0
The coefficient for Occupation_Large Business is 0.0
The coefficient for EducationField_Diploma is 0.0
The coefficient for EducationField_Engineer is 0.0
The coefficient for Gender_Male is 0.0
The coefficient for ExistingProdType_4 is 0.0
The coefficient for ExistingProdType_5 is 0.0
The coefficient for ExistingProdType_6 is 0.0
The coefficient for MaritalStatus_Married is 0.0
The coefficient for MaritalStatus_Single is 0.0
The coefficient for Complaint_0 is 0.0
The coefficient for Complaint_1 is 0.0
The coefficient for Zone_East is 0.0
The coefficient for Zone_South is 0.0
The coefficient for PaymentMethod_Monthly is 0.0
The coefficient for PaymentMethod_Quarterly is 0.0
The coefficient for PaymentMethod_Yearly is 0.0
The coefficient for CustCareScore_5.0 is 0.0
```

c) Interpretation of the model

Gradient Boost:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161
10	Random_Forest1	model_rf	0.9788	0.8563	290651.9734	0.8563
11	Random_Forest2	model_rf1	0.9790	0.8564	290534.3749	0.8564
12	Random_Forest3	model_rf2	0.8374	0.8213	361430.6817	0.8213
13	Random_Forest4	model_rf3	0.8375	0.8213	361478.2054	0.8213
14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350

c) Interpretation of the model

Gradient Boost:

14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350
15	Random_Forest6	model_rf5	0.8588	0.8350	333686.0254	0.8350
16	Gradient_Boost1	model_gb	0.8727	0.8491	305311.5037	0.8491
17	Gradient_Boost2	model_gb1	0.8566	0.8402	323305.7346	0.8402
18	Gradient_Boost3	model_gb2	0.8606	0.8429	317730.7306	0.8429
19	Gradient_Boost4	model_gb3	0.8586	0.8414	320796.7114	0.8414

o

c) Interpretation of the model

Gradient Boost:

Till now best tuned Gradient Boost model is the model_gb2 model with learning rate =0.05 and n_estimators=120.

The error for the model is the least among all the tuned Gradient Boost models till now and has given the highest accuracy for both train and test data and is higher than the Linear regression model or its tuned model of Ridge and Lasso model and also Decision Tree.

The error value is 317730.7306 and the train and test accuracy are 0.8606 and 0.8429 respectively.

a) Model Tuning

Gradient Boost:

There are different parameters to tune the Gradient Boost model such as learning rate, changing number of trees, etc.

I have used GridSearch to fine tune the Gradient Boost model.

The best model which came from it is named as `best_model1` and its parameters are as follows:
`learning_rate=0.05, max_depth=6, random_state=42.`

The best parameters are as:

```
{'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 100}
```

a) Model Tuning

Gradient Boost:

I can see that the model score increases a little. The error decreases for this model to 282701.9276 and the train and ^{test} scores are 0.9228 and 0.8603 respectively.

I can see that this model gives the best accuracy for test data but there is slight overfitting as train score is little higher than the test score.



a) Model Tuning

Gradient Boost:

Therefore, the best tuned Gradient Boost model is `best_model1` and is the best model after `model_gb` model having the highest accuracy for train and test data as 0.9228 and 0.8603 respectively and also the least error of 282701.9276.

There is slight overfitting in the data.

The value for the coefficients can be seen in the next few slides.

- o

a) Model Tuning

Gradient Boost:

```
The coefficient for SumAssured is 0.7639430591559299
The coefficient for MonthlyIncome is 0.07108005105367851
The coefficient for Age is 0.06849981811994652
The coefficient for CustTenure is 0.05981710542023578
The coefficient for ExistingPolicyTenure is 0.008937615539527446
The coefficient for LastMonthCalls is 0.0055540838770366035
The coefficient for Designation_AVP is 0.002544157519737196
The coefficient for NumberOfPolicy is 0.0015296934802632848
The coefficient for ExistingProdType_3 is 0.0015060181627802466
The coefficient for Designation_VP is 0.0012716563076120198
The coefficient for Designation_Senior Manager is 0.0008499831419457634
The coefficient for Designation_Manager is 0.0008404029954792759
The coefficient for EducationField_Under Graduate is 0.0007838234047652462
The coefficient for Designation_Executive is 0.0007690625941490402
The coefficient for Channel_Agent is 0.0005860374530485615
The coefficient for MaritalStatus_Single is 0.0005478359423590817
The coefficient for MaritalStatus_Divorced is 0.0005413859398535892
The coefficient for MaritalStatus_Married is 0.0005282913376806799
The coefficient for EducationField_Diploma is 0.0005045742608296063
The coefficient for Channel_Third Party Partner is 0.0005034724317613388
The coefficient for Channel_Online is 0.00048794975966905363
The coefficient for CustCareScore_2.0 is 0.0004794011988341782
The coefficient for EducationField_Post Graduate is 0.00047031563887960743
The coefficient for CustCareScore_5.0 is 0.00046985138824576597
The coefficient for CustCareScore_3.0 is 0.00046826549265484833
The coefficient for PaymentMethod_Yearly is 0.0004592588874836083
The coefficient for CustCareScore_4.0 is 0.0004474536272903128
The coefficient for CustCareScore_1.0 is 0.00041805492092447633
```

a) Model Tuning

Gradient Boost:

```
The coefficient for Zone_North is -0.0003968248670065427
The coefficient for Gender_Female is 0.00032921678448071814
The coefficient for EducationField_Graduate is 0.0003253918323314906
The coefficient for Zone_East is 0.00032078468875299703
The coefficient for Occupation_Small Business is 0.0003201282400313333
The coefficient for Zone_West is 0.00031008096315496915
The coefficient for ExistingProdType_1 is 0.00030937227475446867
The coefficient for PaymentMethod_Monthly is 0.0002559667445237303
The coefficient for ExistingProdType_4 is 0.0002497933285360478
The coefficient for PaymentMethod_Half Yearly is 0.0002468712279534061
The coefficient for EducationField_Engineer is 0.0002453374133943768
The coefficient for Complaint_1 is 0.00023936557570363953
The coefficient for ExistingProdType_5 is 0.00023126580374913122
The coefficient for Complaint_0 is 0.00022464712534964765
The coefficient for EducationField_MBA is 0.00022277883358553288
The coefficient for Gender_Male is 0.00021788787953570736
The coefficient for ExistingProdType_2 is 0.00019433611237549783
The coefficient for PaymentMethod_Quarterly is 0.00017133460111619334
The coefficient for Occupation_Salaried is 0.00014553646241767578
The coefficient for Occupation_Large Business is 0.00012061561835449331
The coefficient for ExistingProdType_6 is 8.378457029096337e-05
The coefficient for Occupation_Free Lancer is 0.0
The coefficient for Zone_South is 0.0
```

c) Interpretation of the model

Gradient Boost:

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161
10	Random_Forest1	model_rf	0.9788	0.8563	290651.9734	0.8563
11	Random_Forest2	model_rf1	0.9790	0.8564	290534.3749	0.8564
12	Random_Forest3	model_rf2	0.8374	0.8213	361430.6817	0.8213
13	Random_Forest4	model_rf3	0.8375	0.8213	361478.2054	0.8213
14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350

c) Interpretation of the model

Gradient Boost:

14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350
15	Random_Forest6	model_rf5	0.8588	0.8350	333686.0254	0.8350
16	Gradient_Boost1	model_gb	0.8727	0.8491	305311.5037	0.8491
17	Gradient_Boost2	model_gb1	0.8566	0.8402	323305.7346	0.8402
18	Gradient_Boost3	model_gb2	0.8606	0.8429	317730.7306	0.8429
19	Gradient_Boost4	model_gb3	0.8586	0.8414	320796.7114	0.8414
20	Gradient_Boost5	best_model1	0.9228	0.8603	282701.9276	0.8603

c) Interpretation of the model

Gradient Boost:

Till now best tuned Gradient Boost model is the `best_model1` model with `'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 100`.

The error for the model is the least among all the tuned Gradient Boost models till now and has given the highest accuracy for both train and test data and is higher than the Linear regression model or its tuned model of Ridge and Lasso model, Decision Tree and Random Forest.

The error value is 282701.9276 and the train and test accuracy are 0.9228 and 0.8603 respectively.

c) Interpretation of the most optimum model and its implication on the business

	Model	Model_name	Train_score	Test_score	MSE	R2
0	Linear_Regression	model_lr	0.8021	0.7986	407518.4029	0.7986
1	Ridge	ridge	0.8021	0.7986	407507.8720	0.7986
2	Lasso	lasso	0.8021	0.7987	407256.6183	0.7987
3	Decision_Tree1	model_dt	1.0000	0.7175	571543.9624	0.7175
4	Decision_Tree2	model_dt_1	0.7342	0.7195	567341.4877	0.7195
5	Decision_Tree3	model_dt_2	0.8132	0.7913	422189.0950	0.7913
6	Decision_Tree4	model_dt_3	0.8132	0.7913	422189.0950	0.7913
7	Decision_Tree5	model_dt_4	0.8756	0.8143	375658.8632	0.8143
8	Decision_Tree6	model_dt_5	0.9078	0.7949	414981.0028	0.7949
9	Decision_Tree7	model_dt_6	0.8568	0.8161	371980.1317	0.8161
10	Random_Forest1	model_rf	0.9788	0.8563	290651.9734	0.8563
11	Random_Forest2	model_rf1	0.9790	0.8564	290534.3749	0.8564
12	Random_Forest3	model_rf2	0.8374	0.8213	361430.6817	0.8213
13	Random_Forest4	model_rf3	0.8375	0.8213	361478.2054	0.8213
14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350

c) Interpretation of the most optimum model and its implication on the business

14	Random_Forest5	model_rf4	0.8590	0.8350	333725.5135	0.8350
15	Random_Forest6	model_rf5	0.8588	0.8350	333686.0254	0.8350
16	Gradient_Boost1	model_gb	0.8727	0.8491	305311.5037	0.8491
17	Gradient_Boost2	model_gb1	0.8566	0.8402	323305.7346	0.8402
18	Gradient_Boost3	model_gb2	0.8606	0.8429	317730.7306	0.8429
19	Gradient_Boost4	model_gb3	0.8586	0.8414	320796.7114	0.8414
20	Gradient_Boost5	best_model1	0.9228	0.8603	282701.9276	0.8603

c) Interpretation of the most optimum model and its implication on the business

We can see from the table of the score and the error along with the model name that tuning the model has increased the accuracy of the model and decreased the error of the model except that of Linear Regression model where the Ridge and Lasso model didn't have much effect.

Before tuning Gradient Boost was the best model named `model_gb` and after tuning `best_model1` which is Gradient Boost model and overall also it is the best model with the highest accuracy and least error.

o

c) Interpretation of the most optimum model and its implication on the business

The best tuned model is the `best_model1` model which Gradient Boost model with 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 100.

The error for the model is the least among all the tuned Gradient Boost models till now and has given the highest accuracy for both train and test data and is higher than the Linear regression model or its tuned model of Ridge and Lasso model, Decision Tree and Random Forest.

The error value is 282701.9276 and the train and test accuracy are 0.9228 and 0.8603 respectively.



c) Interpretation of the most optimum model and its implication on the business

The top 5 most important feature for this model is as follows:

1. SumAssured with score of 0.7639430591559299
2. MonthlyIncome \cdot with score of 0.07108005105367851
3. Age with score of 0.06849981811994652
4. CustTenure with score of 0.05981710542023578
5. ExistingPolicyTenure with score of 0.008937615539527446.

o

c) Interpretation of the most optimum model and its implication on the business

The most optimum gradient boosting model provides accurate predictions by combining multiple weak learners. It helps identify the key factors impacting the insurance bonus and their interactions. The feature importance scores derived from gradient boosting can guide business decisions related to agent performance evaluation, reward systems, and targeted improvement initiatives.

- o

c) Interpretation of the most optimum model and its implication on the business

If the agent has clients with higher sum assured, higher monthly income and designation of the client is AVP, then⁺ the agent gets higher bonus.

So business can accordingly pay the agent bonus and try to teach the agents who are not performing well to target the important features of this model to get better bonus.

- o

- Since the mean age of customers is 15 years, so we can target married people to opt for company's insurance.
- We can increases the benefits for customers who remain for longer period of time with the company.
- Since, product type 4 is more popular, we can try to sell it to different customers specially the married customers.
- We can focus on finding the issue for low customer score and depending on that conduct training for the agents.
- Since, the agents are preferred by customers as compared to third party, we can hire more agents and train them to get more clients.

- We can target more salaried and small business owners to take up insurance and also focus on why freelancers are not taking insurance or whether we are not targeting them and depending on that focus on onboarding + freelancers and large business owner.
- We can focus more on VP and AVP by giving some personalized insurance policy, as they might bring more revenue as they might be opting for higher premium.
- Focus more on South and East zone and find why there are less customers from here. Get more agents there and get more customers.
- Target married people having children.



THANK YOU

RAMESH PANDEY

ramesh.pandey.1015@gmail.com

9566793109