Telecom Churn Project

Domain Oriented Case Study

Shubham Apurwa

(GCP in Data Science and Artificial Intelligence)

Problem Statement

To analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

TECHNICAL ASPECT OF THE PROJECT

LINEAR REGRESSION : RIDGE AND LASSO

Non Normal Distribution of Residual Terms , We will consider hyperparameter Tuning .

Technical Aspects

1. With respect to technical aspect , we found errors are not normally distributed in case of linear regression . It can be seen in figure given by side.

2. We use Lasso and Ridge Regression , to find out the value of alpha , which helps us to tune the parameters and make decision according to the parameter tuned.

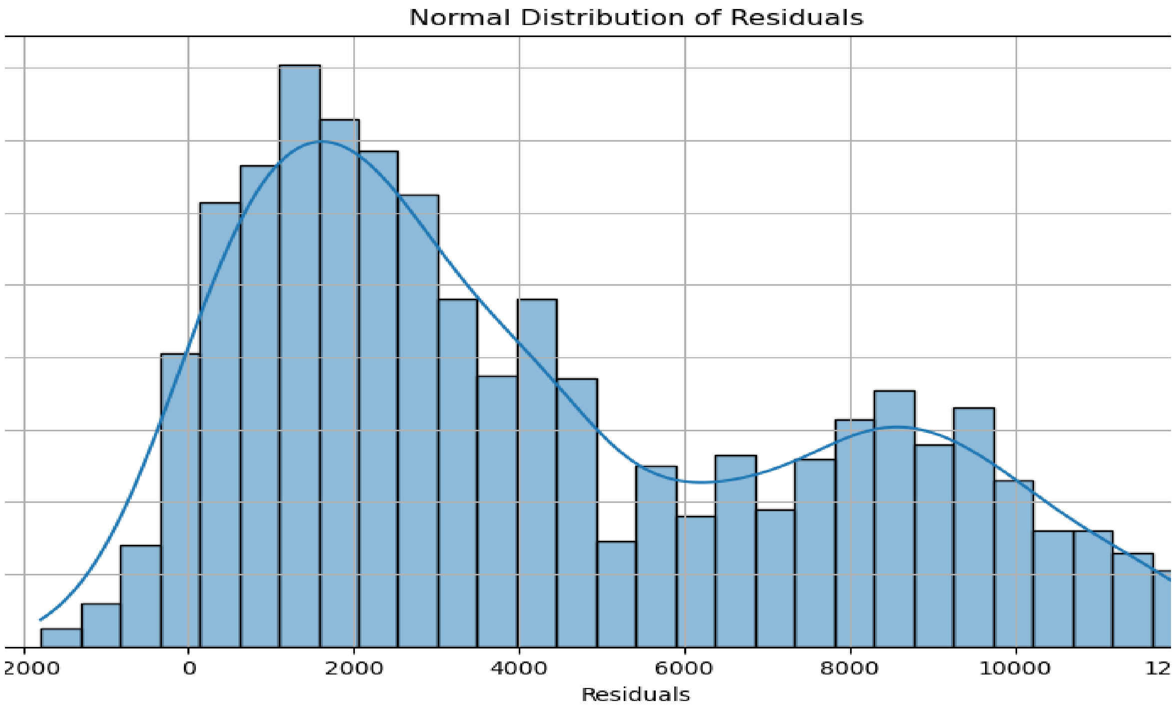
3. As seen in table given , the r2 value for linear regression is not fit .

4. But after hyperparameter tuning , we get the best r2 score value for both Ridge and Lasso .

5. We see RSS(Test) and MSE(Test) values to choose between Ridge and Lasso

6. In the table given , low value of both gives us the best regression model

7. So , we consider Lasso in this case.



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	3.780017e-02	0.004176	0.016662
1	R2 Score (Test)	-1.130888e+09	0.004040	0.021348
2	RSS (Train)	9.228711e+01	95.512095	94.314559
3	RSS (Test)	5.832701e+10	51.367931	50.475240
4	MSE (Train)	1.416265e-01	0.144080	0.143174
5	MSE (Test)	5.437150e+03	0.161355	0.159947

Technical Aspect

1. After hyperparameter tuning , we can see coefficients values of Ridge Regression has been made close to 0.
2. In case of Lasso Regression , we can see that some coefficients of parameters has been made to 0 , so Lasso made important feature selection here.
3. The features selected by Lasso Regression are :

    a) avg_roaming_og_mou

    b) spl_ic_mou_avg

    c) monthly_2g_avg

    d) monthly_3g_avg

    e) sachet_3g_avg

	Linear	Ridge	Lasso
aon	-3.139673	-3.348930e-09	-0.000000
aug_vbc_3g	-0.027506	-1.268490e-08	-0.000000
jul_vbc_3g	0.203508	-1.107186e-08	-0.000000
jun_vbc_3g	0.074775	-1.351520e-08	-0.000000
sep_vbc_3g	-0.004671	-1.825423e-07	-0.000000
avg_2g_usage	0.160540	-1.259699e-08	-0.000000
avg_3g_usage	0.400746	-8.609765e-09	-0.000000
avg_std_og_mou	1.610428	2.015837e-08	0.000000
avg_roaming_ic_mou	0.000057	4.715270e-05	0.000000
avg_roaming_og_mou	0.000688	1.064340e-04	0.000732
offnet_mou_avg	-2.000983	7.820644e-09	0.000000
loc_og_t2t_mou_avg	-0.441916	-5.182869e-08	-0.000000
loc_og_t2f_mou_avg	-0.006842	-4.450013e-07	-0.000000
loc_og_t2c_mou_avg	0.005617	6.500775e-07	0.000000
std_og_t2m_mou_avg	2.977416	2.583264e-08	0.000000
std_og_t2f_mou_avg	-0.000804	-9.166687e-07	-0.000000
isd_og_mou_avg	-0.000048	-1.415824e-06	-0.000000
spl_og_mou_avg	-0.006973	7.187172e-08	0.000000
loc_ic_t2t_mou_avg	0.128869	-5.499295e-08	-0.000000
loc_ic_t2f_mou_avg	-0.015307	-1.983241e-07	-0.000000
std_ic_t2t_mou_avg	-0.027650	6.469302e-08	0.000000
std_ic_t2f_mou_avg	-0.005621	-1.316362e-06	-0.000000
spl_ic_mou_avg	-0.000279	-1.481805e-05	-0.000173
isd_ic_mou_avg	-0.020542	-1.739297e-07	-0.000000
max_rech_amt_avg	-0.617211	-6.279696e-08	-0.000000
monthly_2g_avg	-0.001024	-8.996271e-06	-0.000968
sachet_2g_avg	-0.001912	-9.056875e-07	-0.000000
monthly_3g_avg	-0.000533	-6.905096e-06	-0.000597
sachet_3g_avg	-0.000554	-8.162506e-06	-0.000393
total_rech_num_avg	-0.031913	2.400647e-07	0.000000

LOGISTIC REGRESSION

We are considering building Logistic Regression Model , because we have binary classification . i.e. the customer will churn or not

Technical Aspect

1. While building the model with Logistic Regression , we used the RFE (Recursive Feature Elimination) method to select top predictor variable and then do manual elimination
2. After selecting top 15 predictor variable and removing the insignificant variable using p-values , we found out the following top 5 parameters :
   a) avg_std_og_mou
   b) loc_og_t2t_mou_avg
   c) loc_og_t2f_mou_avg
   d) std_ic_t2f_mou_avg
   e) max_rech_amt_avg

Dep. Variable:	Churn	**No. Observations:**	4601
Model:	GLM	**Df Residuals:**	4595
Model Family:	Binomial	**Df Model:**	5
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-412.69
Date:	Sat, 14 Sep 2024	**Deviance:**	825.39
Time:	20:43:31	**Pearson chi2:**	4.12e+03
No. Iterations:	10	**Pseudo R-squ. (CS):**	0.02636
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-33.6757	42.513	-0.792	0.428	-116.999	49.648
avg_std_og_mou	91.5155	32.687	2.800	0.005	27.450	155.581
loc_og_t2t_mou_avg	-55.5194	20.066	-2.767	0.006	-94.849	-16.190
loc_og_t2f_mou_avg	-9.9304	3.880	-2.559	0.010	-17.535	-2.326
std_ic_t2f_mou_avg	-3.3461	1.493	-2.242	0.025	-6.271	-0.421
max_rech_amt_avg	-26.3822	10.324	-2.555	0.011	-46.617	-6.147

Technical Aspect

1. While selecting the 5 parameters during RFE method , we saw only the significance of the parameters
2. After considering the VIF values and elimination of the high VIF valued parameters , we get the following two variables
3. The two variables are :

   a) avg_std_og_mou

   b) std_ic_t2f_mou_avg

Dep. Variable:	Churn	**No. Observations:**	4601
Model:	GLM	**Df Residuals:**	4598
Model Family:	Binomial	**Df Model:**	2
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-434.26
Date:	Sat, 14 Sep 2024	**Deviance:**	868.52
Time:	20:44:26	**Pearson chi2:**	3.97e+03
No. Iterations:	10	**Pseudo R-squ. (CS):**	0.01719
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	152.9720	26.075	5.867	0.000	101.866	204.078
avg_std_og_mou	173.9858	28.314	6.145	0.000	118.490	229.481
std_ic_t2f_mou_avg	-4.6128	1.622	-2.844	0.004	-7.792	-1.434

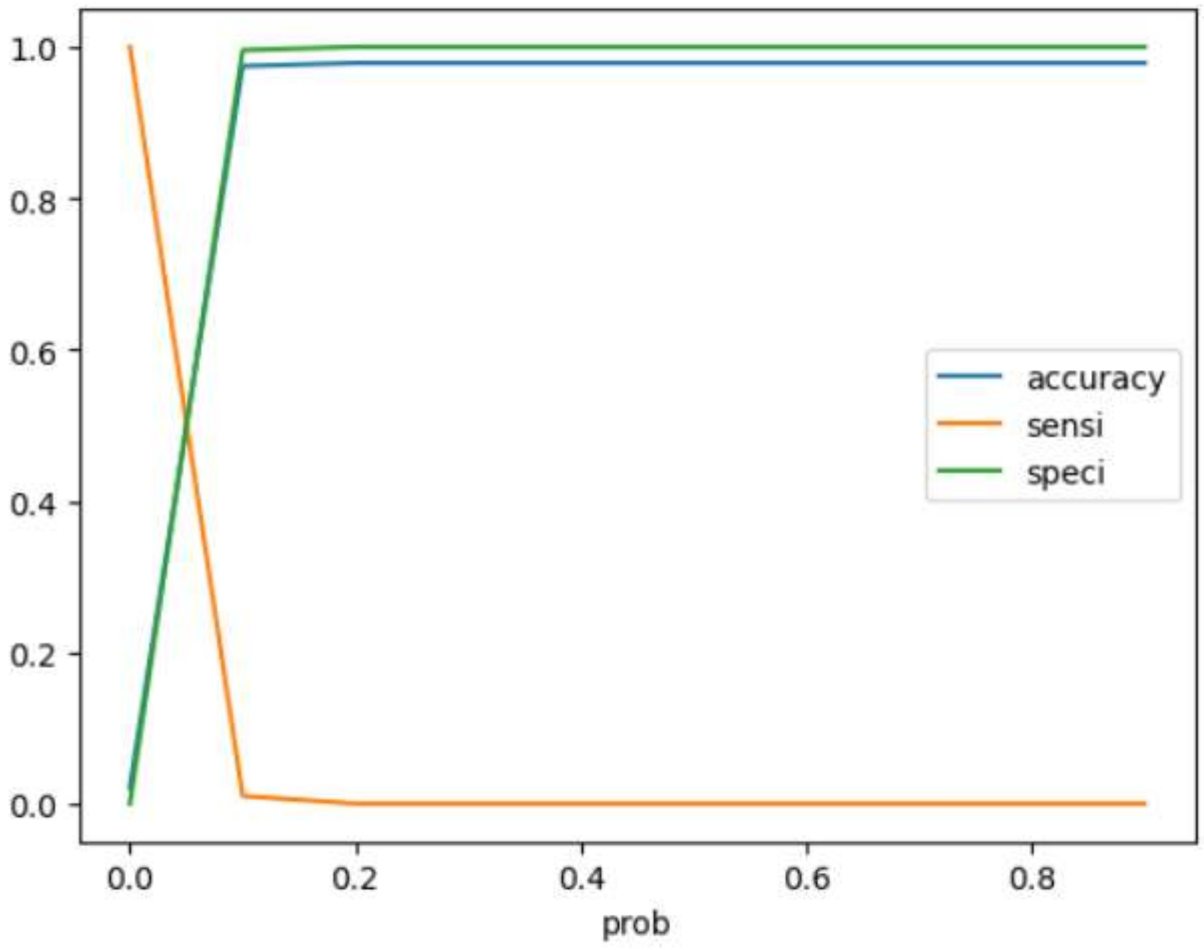
	Features	**VIF**
0	avg_std_og_mou	1.74
1	std_ic_t2f_mou_avg	1.74

FINDING OPTIMAL CUT-OFF

Technical Aspect

1. As in previous models , we considered probability of a customer to be churn as greater than 0.5 .
2. Due to which , we have overfitting of the model .
3. To make fit model and more predictable , we will find optimal cut off .
4. After plotting the graph , we found that cut off where , accuracy , sensitivity and specificity of the model makes fit .
5. The cut off is 0.06

Note : Here the cut off is low because we have data imbalance .



BUSINESS ASPECT OF THE PROJECT

Business Aspect

1. With respect to the Business Aspect , if a customer will churn or not churn , we can consider parameters containing :

INCOMING (LOCAL & ROAMING)

OUTGOING (LOCAL & ROAMING)

2G USAGE

3G USAGE

TOTAL NUMBER OF RECHAGRGE DONE IN A MONTH

MAXIMUM AMOUNT OF RECHARGE

Business Aspect

- The telecom company must see whether the customer is using Offnet Services like Calls , Messages more OR Onnet Services Like Whatsapp , Instagram , Facebook etc.
- This can be identified by seeing the Recharge history of the customer mobile number , which Package he is selecting , whether it is call and messages pack OR Data pack .
- The telecom company must see his number of recharges and amount of recharge he/she is doing in a month
- Based on above criteria , the telecom company must launch effective Calling package or Data package based on customer consumption history .
- The telecom company must also take into consideration whether the competitor telecom company tariff plans , to make customer to not to switch to other network services by doing surveys and feedback and suggestions .
- The above consideration and action plan by the telecom company will make customer not to churn.