Case Study – Telecom Churn

By:

Aishwarya Kumar Sharma

Shahul Hameed

Ashrit Gaikwad

Steps Performed

Data Pre-processing

Feature Engineering

Data Splitting

Data Balancing

Model Selection

Model Training

Model Evaluation

Feature Importance Analysis

Interpretability and Explainability

Actionable Insights & Reporting

Note – *All steps with insights , decisions and actions taken are explained in markdown or comments of Python – Jupyter Notebook*

Data Preprocessing

High Value Customer – Identified a total of 30011 customers as high valued

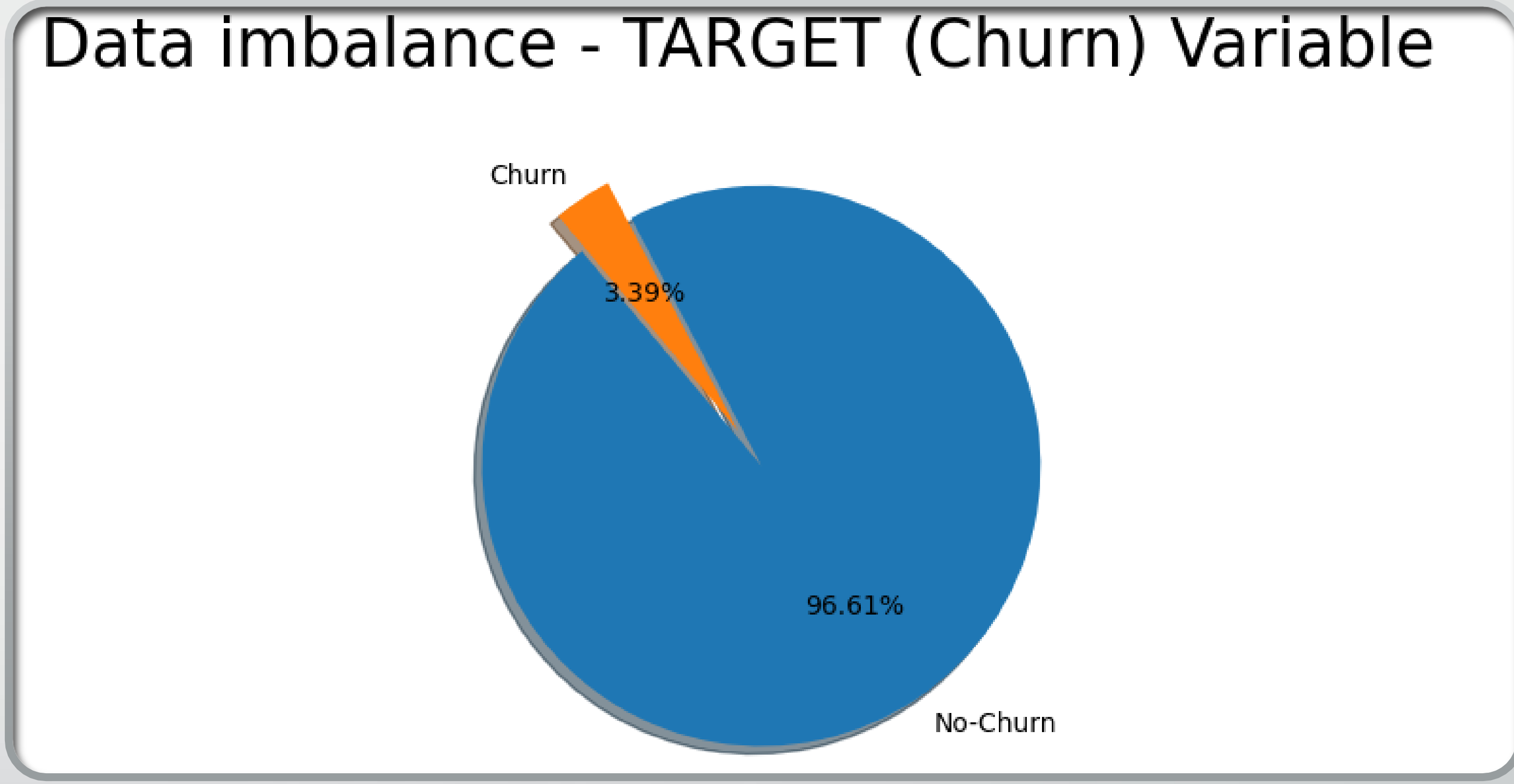
Handling missing value – Missing or NaN value in data has been handled during EDA. Around 7% of data was pruned. However, the remaining dataset still provides us with a sufficient number of records to conduct our analysis effectively.

Outlier Treatment – around 1% of outliers data has been filtered.

Churn Tagging

INFERENCE :

We have identified 3.4% of available data as ‘Churn Customer’ which is a class imbalance issue in our dataset, with a very low percentage of churned customers. We will address this class imbalance later in our analysis.



Feature Engineering

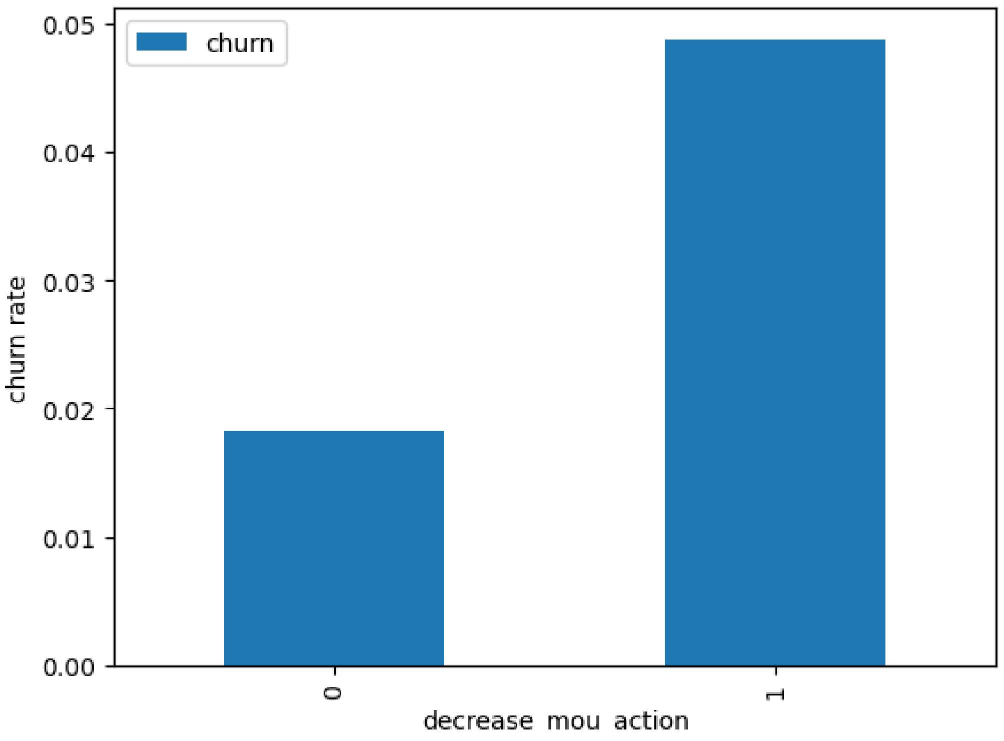
We have derived below new attributes to help to further analysis.

- **‘decrease_mou_action'** to Identify Decreased Usage in the Action Phase
- **decrease_rech_num_action’** to indicate if the number of recharge of the customer has decreased in the action phase than the good phase.
- **decrease_rech_amt_action’** to indicate if the amount of recharge of the customer has decreased in the action phase than the good phase.
- **‘decrease_arpu_action'** to indicate if the average revenue per customer has decreased in the action phase than the good phase.
- **‘decrease_vbc_action'** to indicate if the volume based cost of the customer has decreased in the action phase than the good phase.

EDA & Insights

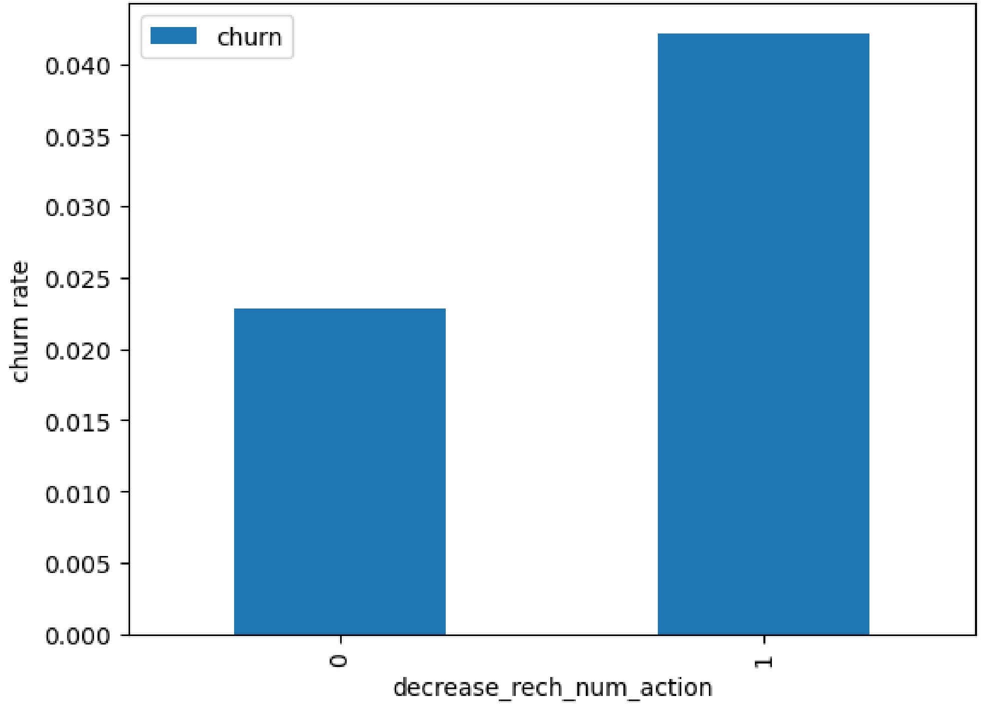
Mixture of Bar chart, scatterplot & Distplots are used

Churn Rate - Customer Decreased MOU in Action Month



Insights: The churn rate is higher for customers whose minutes of usage (MOU) decreased in the action phase compared to the good phase.

Churn Rate - Decrease in Number of Recharges

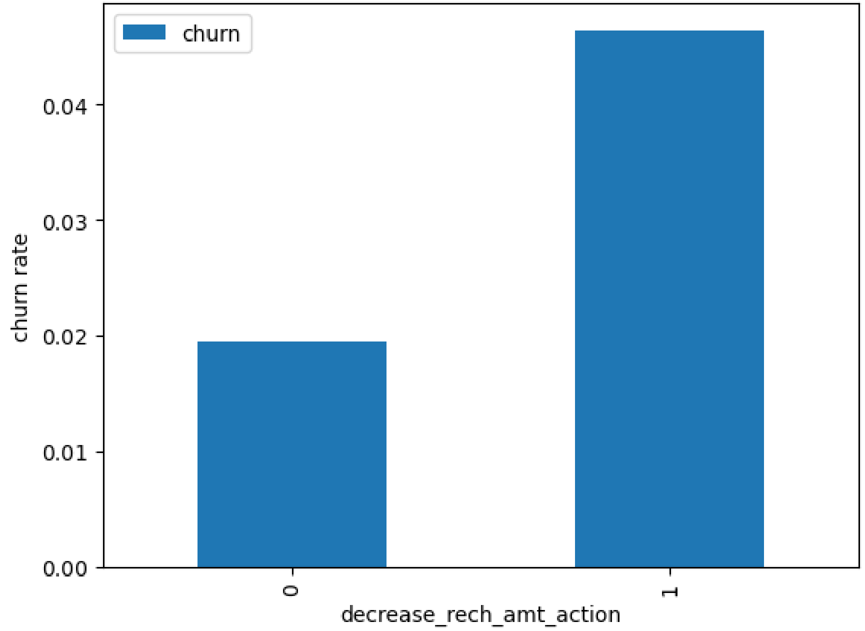


Insights: The churn rate is higher for customers whose number of recharges in the action phase is lower than the number in the good phase.

Churn Rate - Decrease their recharge amount in the action month

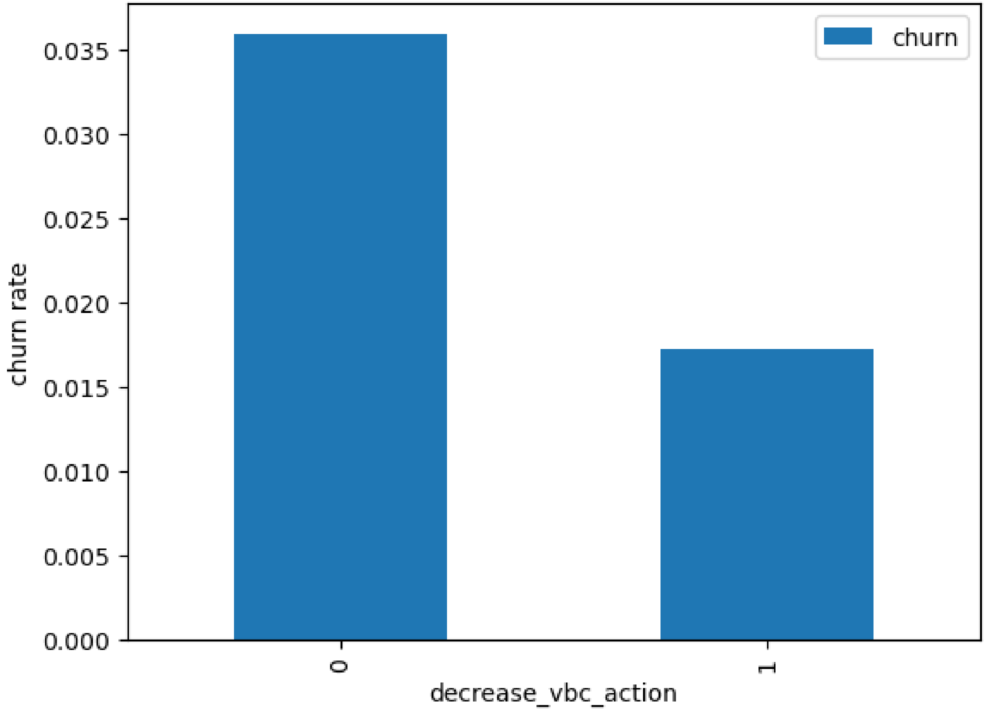
Insights: Decrease in recharge amount during the action phase has higher churn rate compared to stable or increased recharge amounts during the good phase.

This highlights the significance of monitoring and encouraging customers to maintain or boost their recharge amounts to reduce churn risk.



Churn Rate – Decrease in customer volume based cost in action month

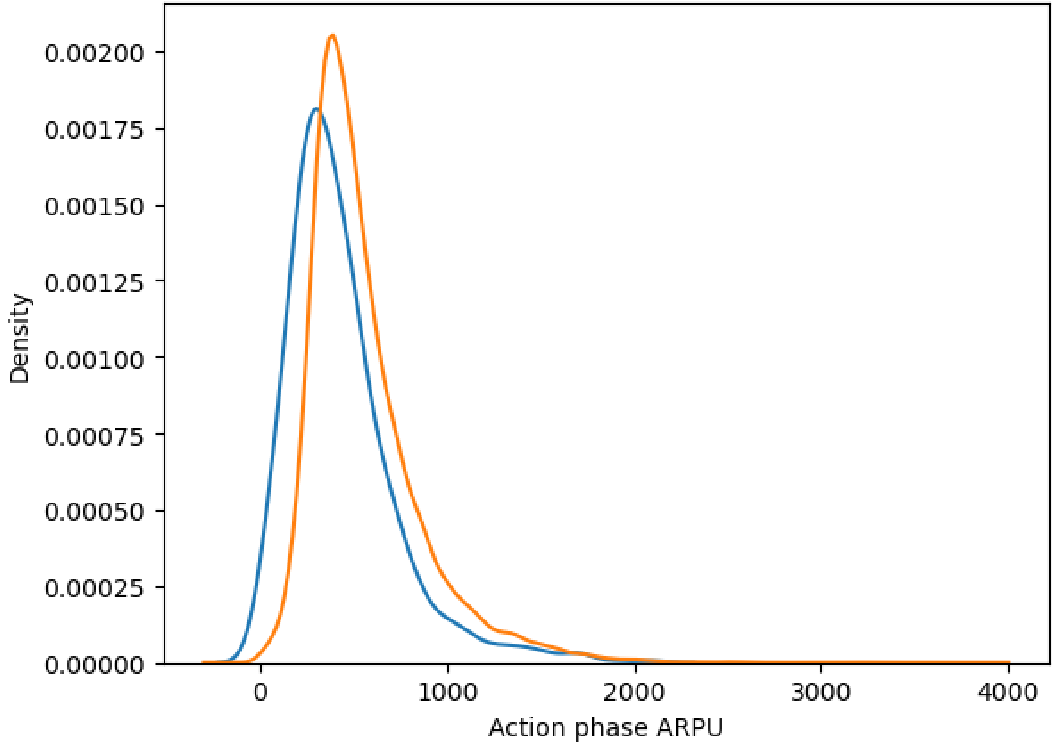
Insights: Higher churn rate is observed among customers whose volume-based cost increased during the action month. This suggests that customers may be recharging less during this phase, indicating a potential risk of churn



Churn Rate - Average revenue per customer in the action phase

Analysis

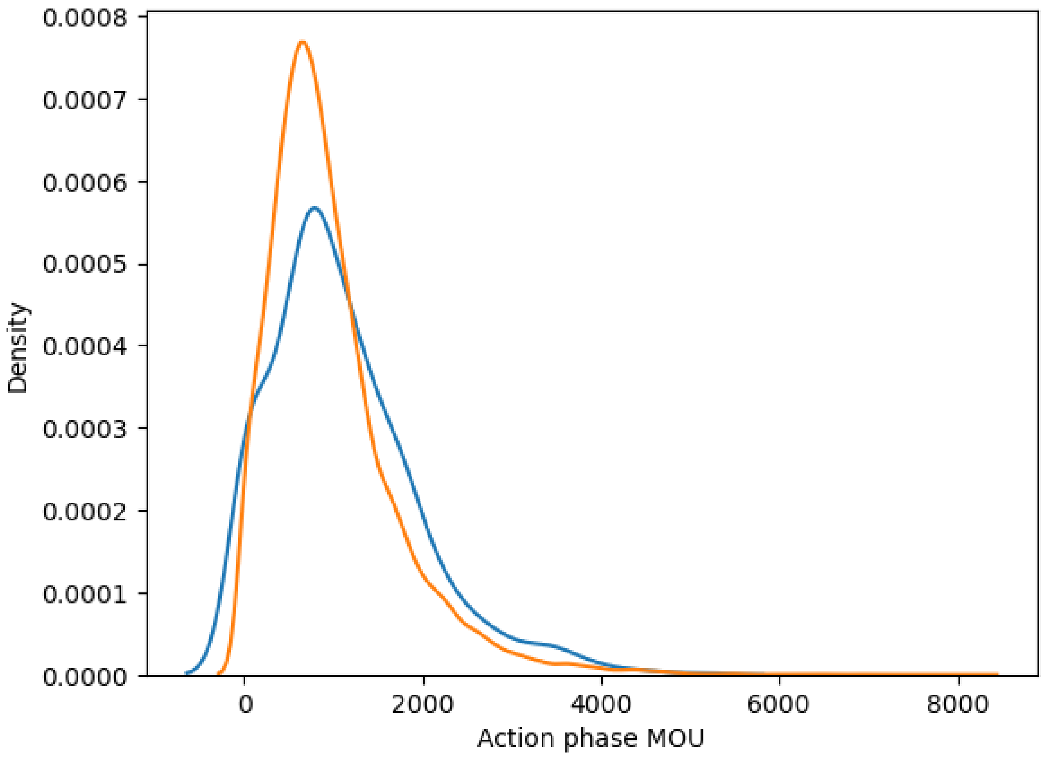
- For churned customers, the Average Revenue Per User (ARPU) is predominantly concentrated in the range of 0 to 900. This indicates that a significant portion of churned customers tends to have lower ARPU.
- In contrast, non-churned customers have their ARPU distribution concentrated in the range of 0 to 1000. This suggests that a majority of non-churned customers tend to have slightly higher ARPU compared to churned customers.
- Notably, higher ARPU customers appear to be less likely to churn, as indicated by the lower density of churned customers in the higher ARPU ranges.



Churn Rate - Minutes of usage MOU in the action phase

Analysis

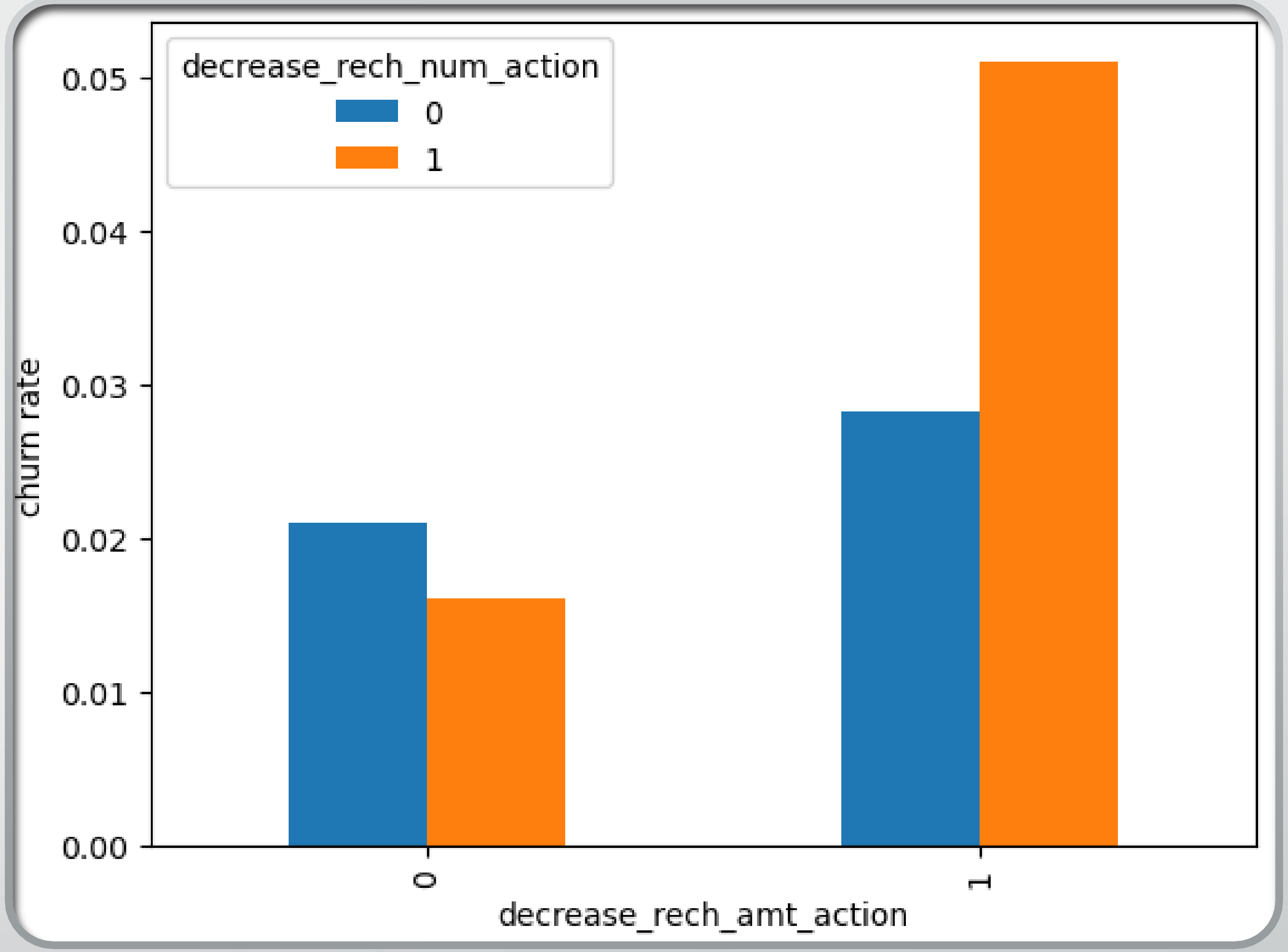
- Among churned customers, the distribution of Minutes of Usage (MOU) is prominently centered in the range of 0 to 2500. This suggests that a substantial portion of churned customers tends to have lower MOU.
- Conversely, lower churn probability is observed for customers with higher MOU, as indicated by the decreased density of churned customers in the higher MOU ranges.
- This finding marks the trend that customers with more significant MOU are less likely to churn, potentially due to their higher engagement and usage of telecom services.



Churn rate - decreasing recharge amount and number of recharge -action phase

Analysis

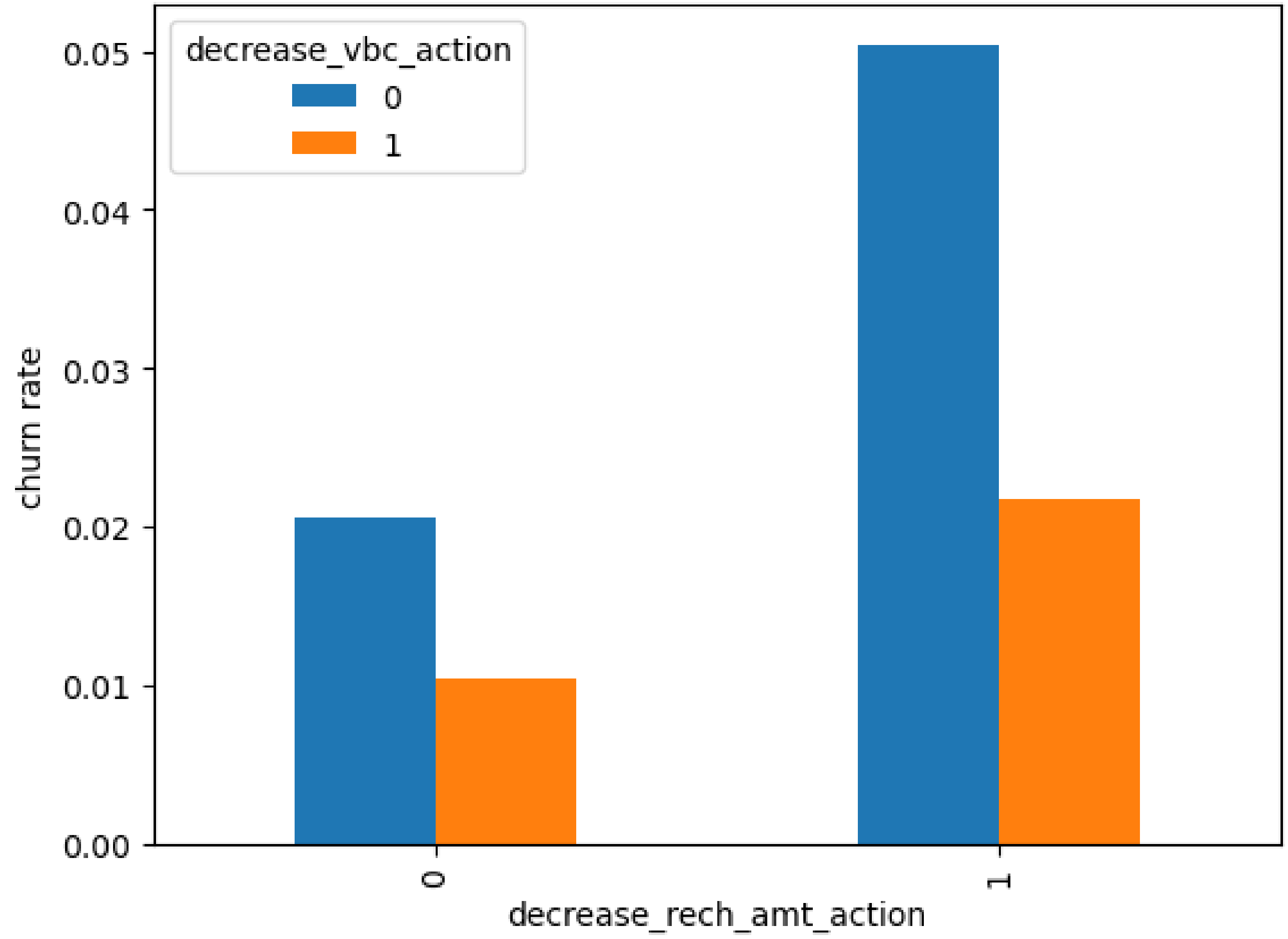
The above plot illustrates that the churn rate is higher for customers whose recharge amount and the number of recharges have both decreased in the action phase compared to the good phase. This trend suggests that a reduction in both recharge amount and frequency of recharges during the action phase is associated with a higher likelihood of churn.



Churn rate - decreasing recharge amount and volume based cost - action phase

Analysis

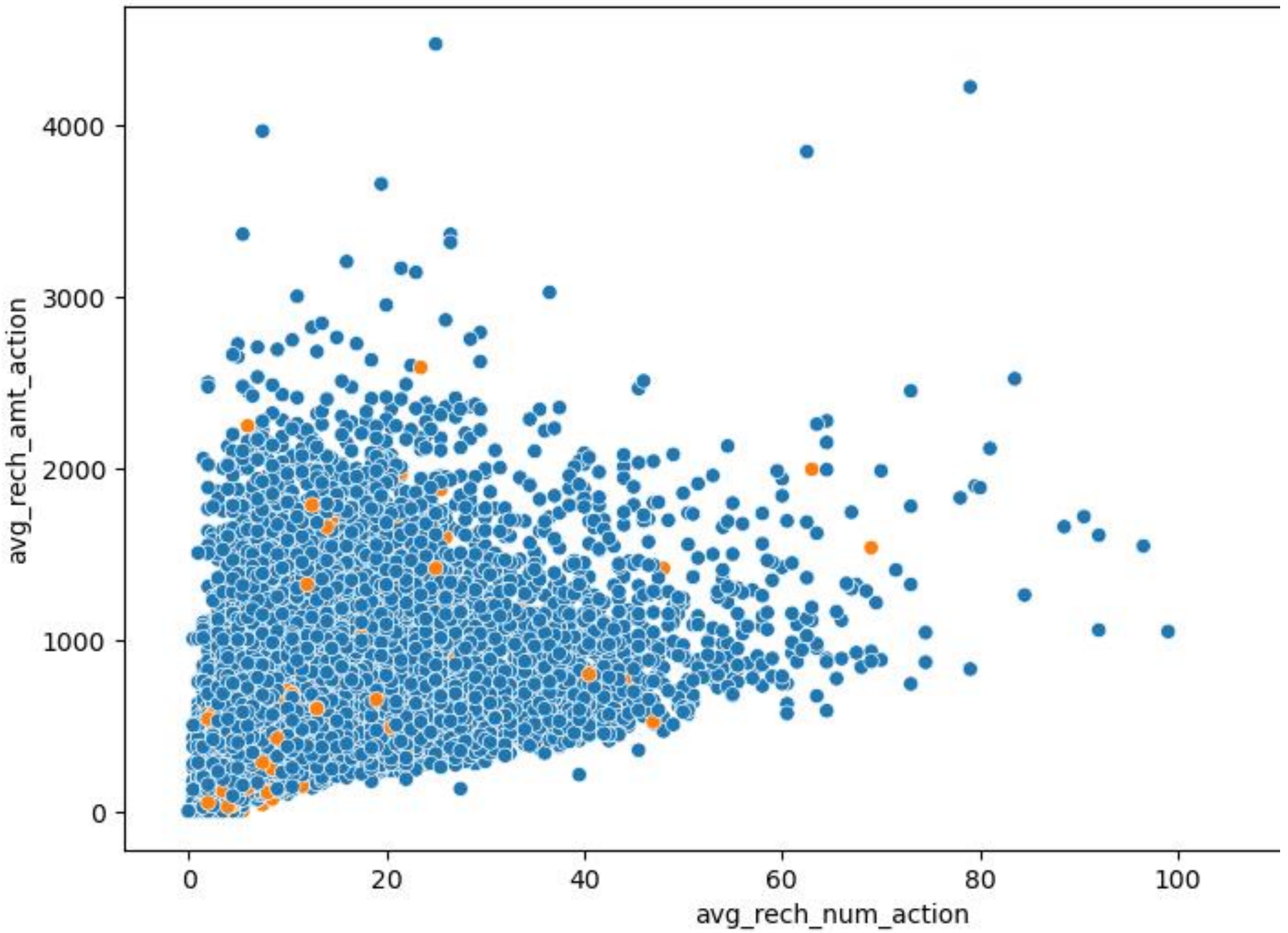
We observe that the churn rate is higher for customers whose recharge amount has decreased in the action month, coinciding with an increase in the volume-based cost. This finding reinforces the notion that a decrease in recharge amount coupled with an increase in cost during the action phase is indicative of a higher likelihood of churn.



Recharge amount and number of recharge in action month

Analysis

The pattern observed in the above plot reveals a strong positive correlation between the number of recharges and the recharge amount. In other words, as the number of recharges increases, so does the total recharge amount.



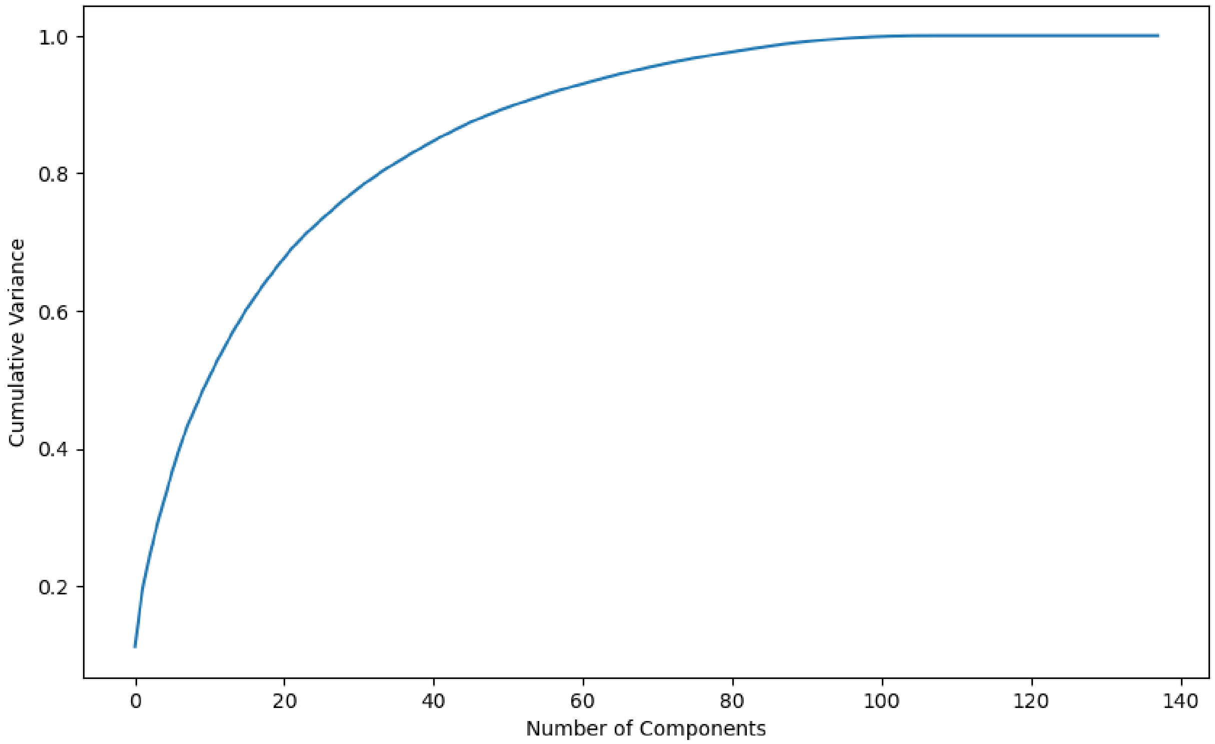
Data Balancing & feature Scaling

- We are generating synthetic samples by applying up-sampling using the SMOTE (Synthetic Minority Oversampling Technique) method.
- We used StandardScaler from sklearn for feature scaling of numerical attributes

Model with PCA

Inference:

We can observe that using `60 components` explains nearly 90% of the data variance. Thus, we will proceed with PCA using 60 components.



Data Modelling

We have created below Machine Learning Models

Models With Dimensionality reduction method (PCA):

- *Logistic Regression*
- *Classification Model*
- *Decision Tree*
- *Random Forest*

With RFE:

- *Logistic Regression (fine tuned with p-Value & VIF measure)*

Logistic Regression Model with PCA

The hyperparameter C in Logistic Regression represents the inverse of the regularization strength. Higher values of C result in less regularization.

Optimal value of C that balances the trade-off between fitting the training data well and preventing overfitting.

We have done Hyperparameter tuning to determine Optimal C Value as ‘100’. The highest test sensitivity is 0.8978916608693863 at C = 100

Built the Logistic regression model with Optimal C-Vale(100) & 60 component using PCA features.

Model Summary:

Train Set

- Accuracy: 0.86
- Sensitivity (Recall): 0.89
- Specificity: 0.83

Test Set

- Accuracy: 0.83
- Sensitivity (Recall): 0.81
- Specificity: 0.83

Overall, the model is performing well in the test set, demonstrating that it has effectively generalized from the training set.

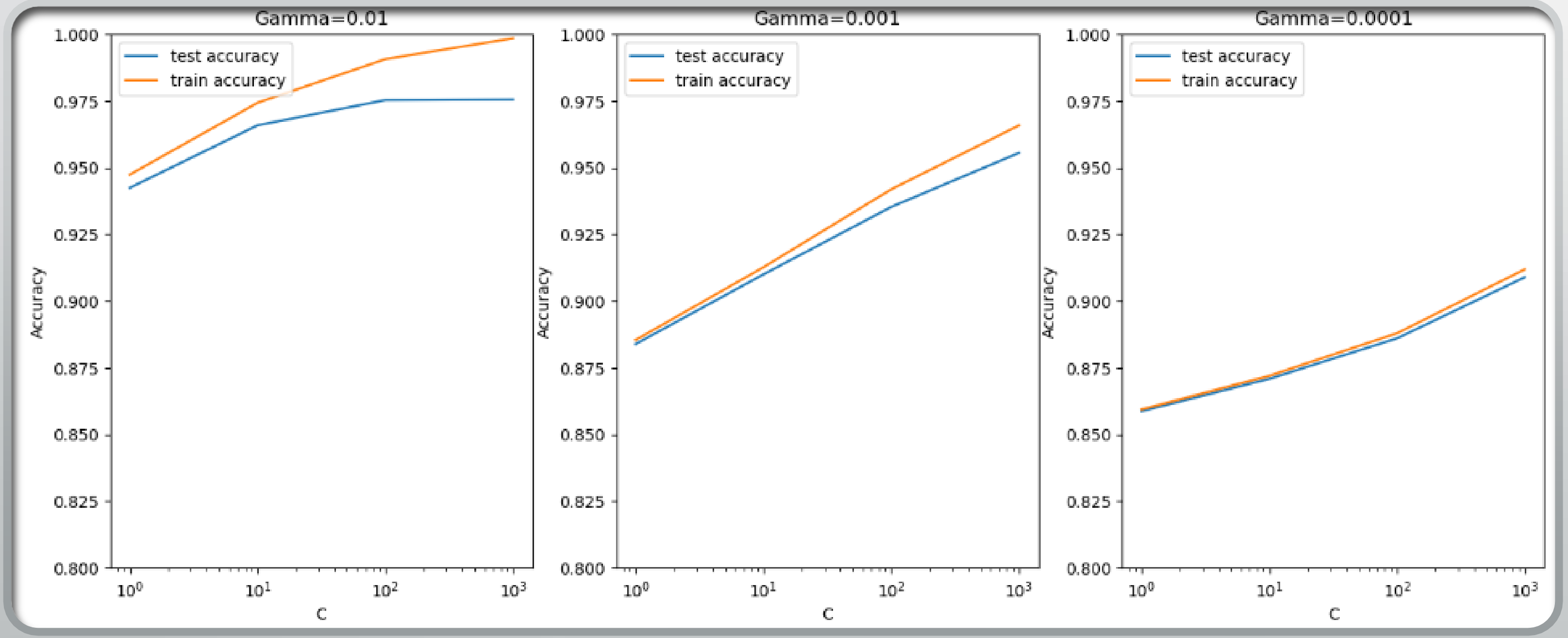
Support Vector Classification(SVC) with PCA

Hyperparameter Tuning

- **C**: Regularization Parameter
- **gamma**: Controls Non-linear Classifications

These hyperparameters play crucial roles in shaping the model's performance and behavior.

The best test score is 0.9754959911159373 corresponding to hyperparameters {'C': 1000, 'gamma': 0.01}



Hyperparameter Tuning & Insights with PCA for Classification Model (SCV)

Insights from Hyperparameter Tuning

Here are the insights we've gained from hyperparameter tuning:

- **Effect of Gamma:** As shown in the plot, a higher value of gamma tends to lead to overfitting the model. With the lowest value of gamma (0.0001), we observe that the train and test accuracy are nearly the same.
- **Optimal C Value:** At C=100, we achieve a good accuracy, and both the train and test scores are comparable. This suggests that a simpler, more linear model with gamma=0.0001 can perform as well as the model with higher gamma values.
- **Tradeoff Consideration:** While sklearn suggests optimal scores (gamma=0.01, C=1000), we argue that choosing a simpler model with gamma=0.0001 is a valid option. This tradeoff involves balancing high gamma (high non-linearity) and average C value or low gamma (less non-linearity) and a high C value.
- **Model Simplicity:** We prioritize model simplicity and choose gamma=0.0001 and a high C=100 to achieve a comparable average test accuracy of around 90%. This choice leans towards a model with less non-linearity.

By making this choice, we aim to balance model complexity while maintaining high Sensitivity/Recall accuracy.

Hyperparameter Tuning & Insights with PCA for Classification Model (SCV)

Built the SVC model with optimal scores (gamma=0.01, C=1000) using PCA features.

Model summary

Train set

- Accuracy = 0.89
- Sensitivity = 0.92
- Specificity = 0.85

Test set

- Accuracy = 0.85
- Sensitivity = 0.81
- Specificity = 0.85

Overall, the model is performing well in the test set, demonstrating that it has effectively generalized from the training set.

Decision tree with PCA

The hyperparameter like max_depth, min_samples_leaf, min_samples_split of decision tree are determined using GridSearchCV HyperParameter tuning approach.

Best sensitivity:- 0.9004900816802801 DecisionTreeClassifier (max_depth=10, min_samples_leaf=50, min_samples_split=50)

Model Summary

Train Set:

- Accuracy = 0.90
- Sensitivity = 0.91
- Specificity = 0.88

Test Set:

- Accuracy = 0.86
- Sensitivity = 0.70
- Specificity = 0.87

In this summary, we can see that while the model maintains high accuracy and specificity, there is a slight decrease in sensitivity when applied to the test set. This suggests room for improvement in correctly identifying churned customers

Random forest with PCA

The hyperparameter like max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators of Random Forest are determined using GridSearchCV HyperParameter tuning approach.

We can get accuracy of 0.8451808791023977 using {'max_depth': 5, 'max_features': 20, 'min_samples_leaf': 50, 'min_samples_split': 100, 'n_estimators': 100}

Model Summary

Train Set:

- Accuracy = 0.84
- Sensitivity = 0.88
- Specificity = 0.80

Test Set:

- Accuracy = 0.80
- Sensitivity = 0.75
- Specificity = 0.80

While the sensitivity decreased when evaluating the model on the test set, it still maintains good levels of accuracy and specificity.

Conclusion with PCA models

After exploring the below models, it's evident that both classic Logistic Regression and SVM models perform well in achieving the best sensitivity, which was our ultimate goal. Both models achieved a sensitivity of approximately 81% while maintaining a good accuracy of around 85%.

- Logistic Regression
- Classification Model
- Decision Tree
- Random Forest

Logistic Regression Model without PCA

Logistic Regression model has been built on train data and coarse tuning has been done with RFE (15 features)& manual method by reviewing the p-value & VIF measures iteratively creating multiple models.

Cut-off value:

Final Model Summary:

Train Set

- Accuracy: 0.86
- Sensitivity (Recall): 0.89
- Specificity: 0.83

Test Set

- Accuracy: 0.83
- Sensitivity (Recall): 0.81
- Specificity: 0.83

Overall, the model is performing well in the test set, demonstrating that it has effectively generalized from the training set.

Logistic Regression Model without PCA

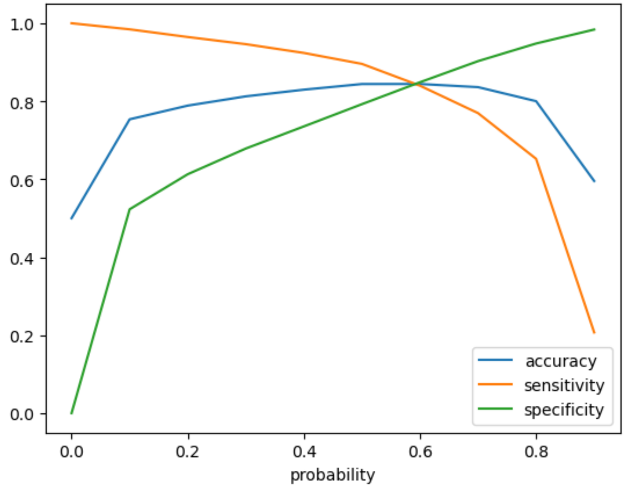
Logistic Regression model has been built on train data and coarse tuning has been done with RFE (15 features)& manual method by reviewing the p-value & VIF measures iteratively creating multiple models.

***Optimal Probability Cut-off Point using graph*:**

- Accuracy: The curve shows that accuracy becomes stable at around 0.6 on the probability scale.
- Sensitivity: Sensitivity decreases as the probability threshold increases.
- Specificity: Specificity increases as the probability threshold increases.

At the point where these three parameters intersect, which is at a probability threshold of 0.6, there is a balance between sensitivity and specificity, resulting in good accuracy.

While the curve suggests taking 0.6 as the optimal probability cutoff, we have chosen ***0.5*** to prioritize higher sensitivity, which aligns with our primary goal.



Logistic Regression Model without PCA

We can see the area of the ROC curve is closer to 1, which is the Gini of the model.

Model summary

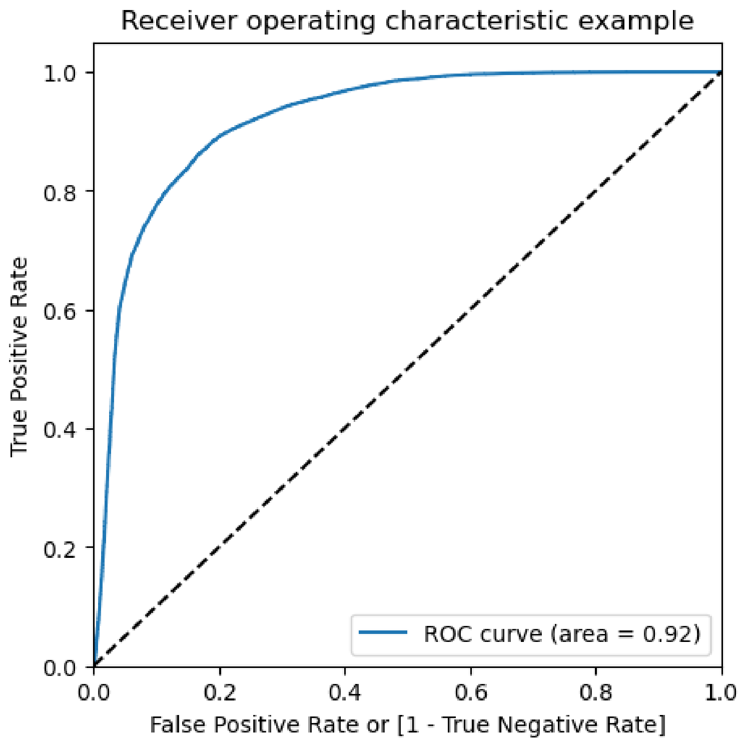
Train set

- Accuracy = 0.84
- Sensitivity = 0.81
- Specificity = 0.83

Test set

- Accuracy = 0.78
- Sensitivity = 0.82
- Specificity = 0.78

Overall, the model is performing well in the test set, what it had learnt from the train set.



Conclusion without PCA - Logistic Regression

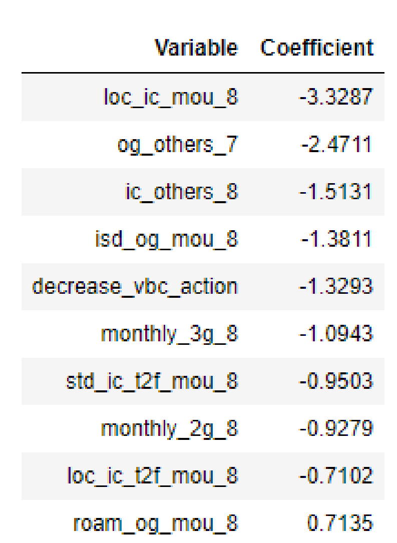
The logistic model without PCA exhibits strong sensitivity and accuracy, which are on par with models using PCA. Therefore, opting for a simpler model like logistic regression without PCA is a valid choice. This model effectively highlights the key predictor variables and their significance, making it a valuable tool for decision-making regarding potential churned customers. Consequently, this model offers enhanced interpretability and relevance for business insights.

Business recommendations

Top Predictors

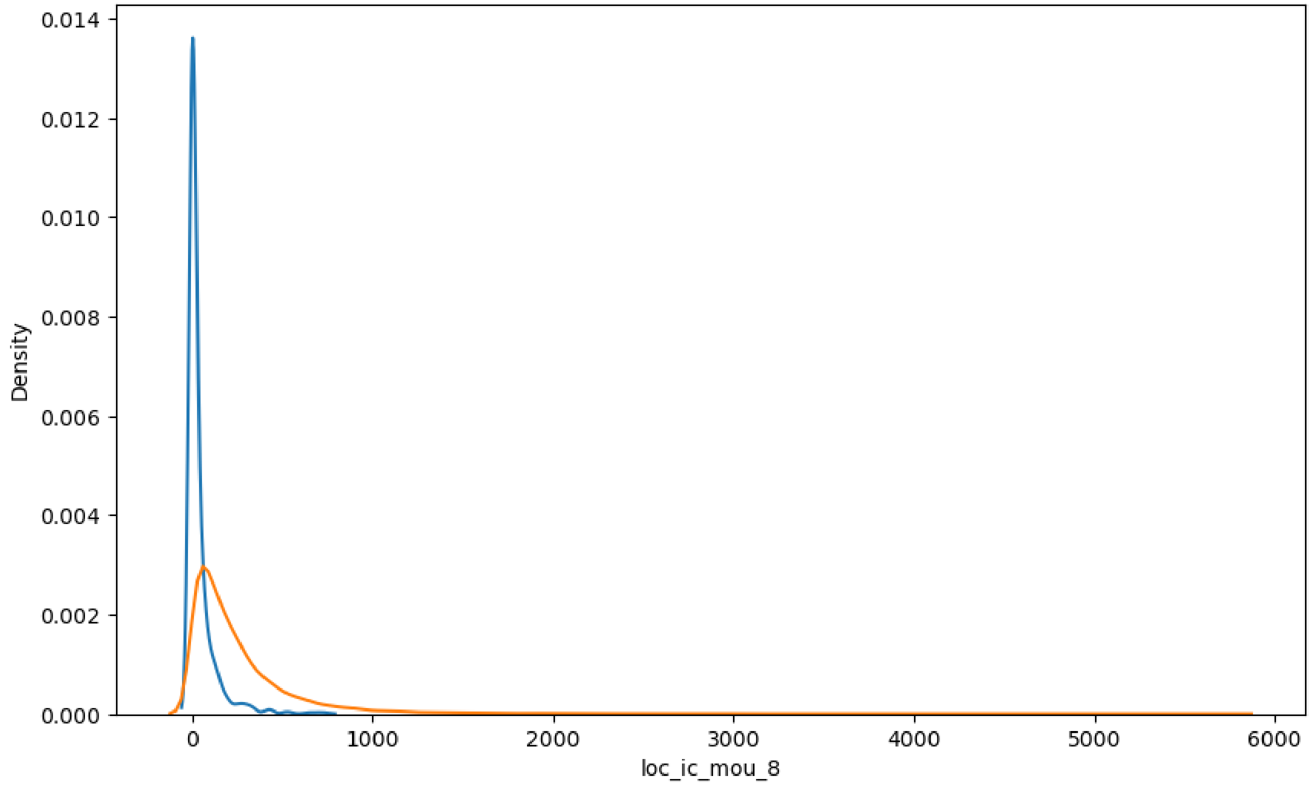
The logistic regression model has identified several key variables that strongly influence churn probability. Below are some of the top predictors along with their coefficients:

- Most of the top variables exhibit negative coefficients, indicating an inverse correlation with churn probability. For instance:
- Customers with lower local incoming minutes of usage (loc_ic_mou_8) in August are more likely to churn.
- Decreased outgoing charges to other operators (og_others_7) in July and decreased incoming charges from other operators (ic_others_8) in August are indicative of higher churn probability.
- An increase in value-based cost (decrease_vbc_action) during the action phase raises the likelihood of churn.



Variable	Coefficient
loc_ic_mou_8	-3.3287
og_others_7	-2.4711
ic_others_8	-1.5131
isd_og_mou_8	-1.3811
decrease_vbc_action	-1.3293
monthly_3g_8	-1.0943
std_ic_t2f_mou_8	-0.9503
monthly_2g_8	-0.9279
loc_ic_t2f_mou_8	-0.7102
roam_og_mou_8	0.7135

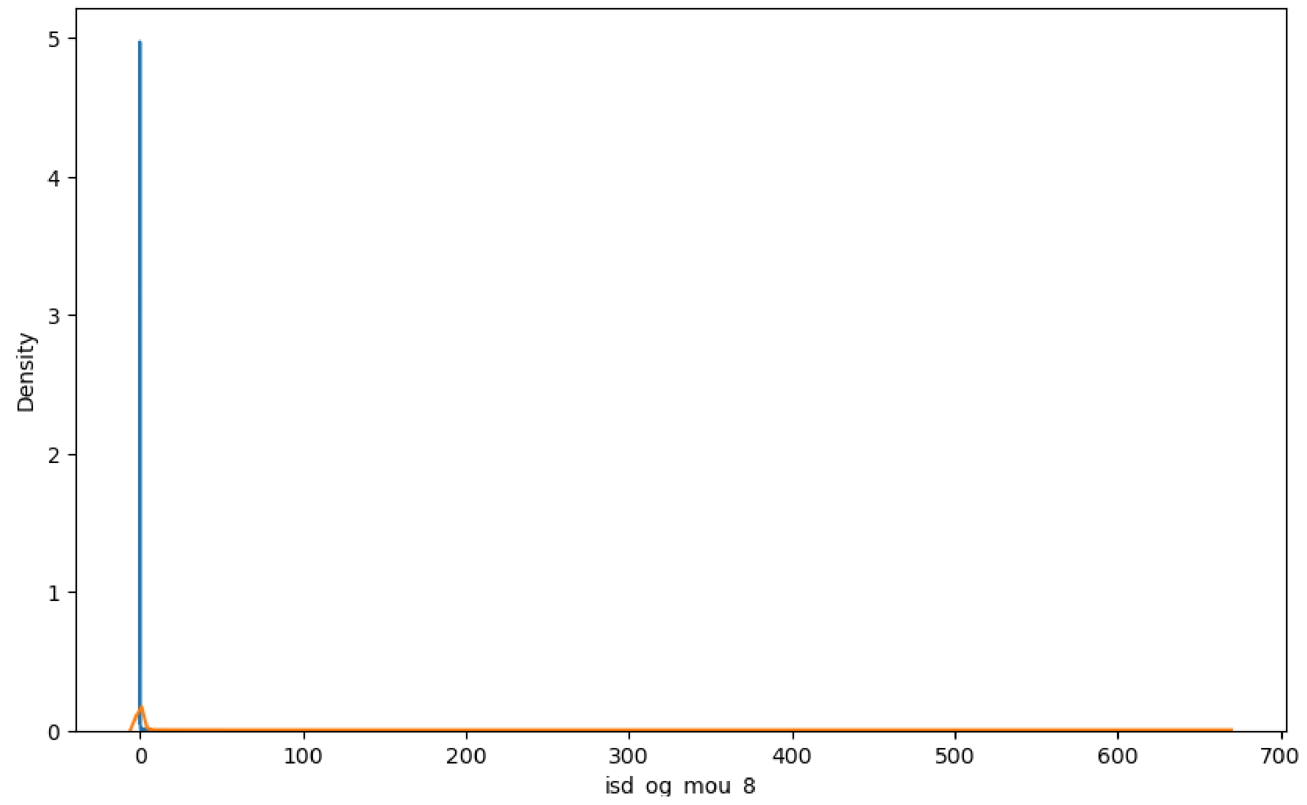
Important Predictor: Minutes of Usage



Insights:

We can observe that, for churned customers, the minutes of usage in August are predominantly skewed towards lower values compared to non-churn customers.

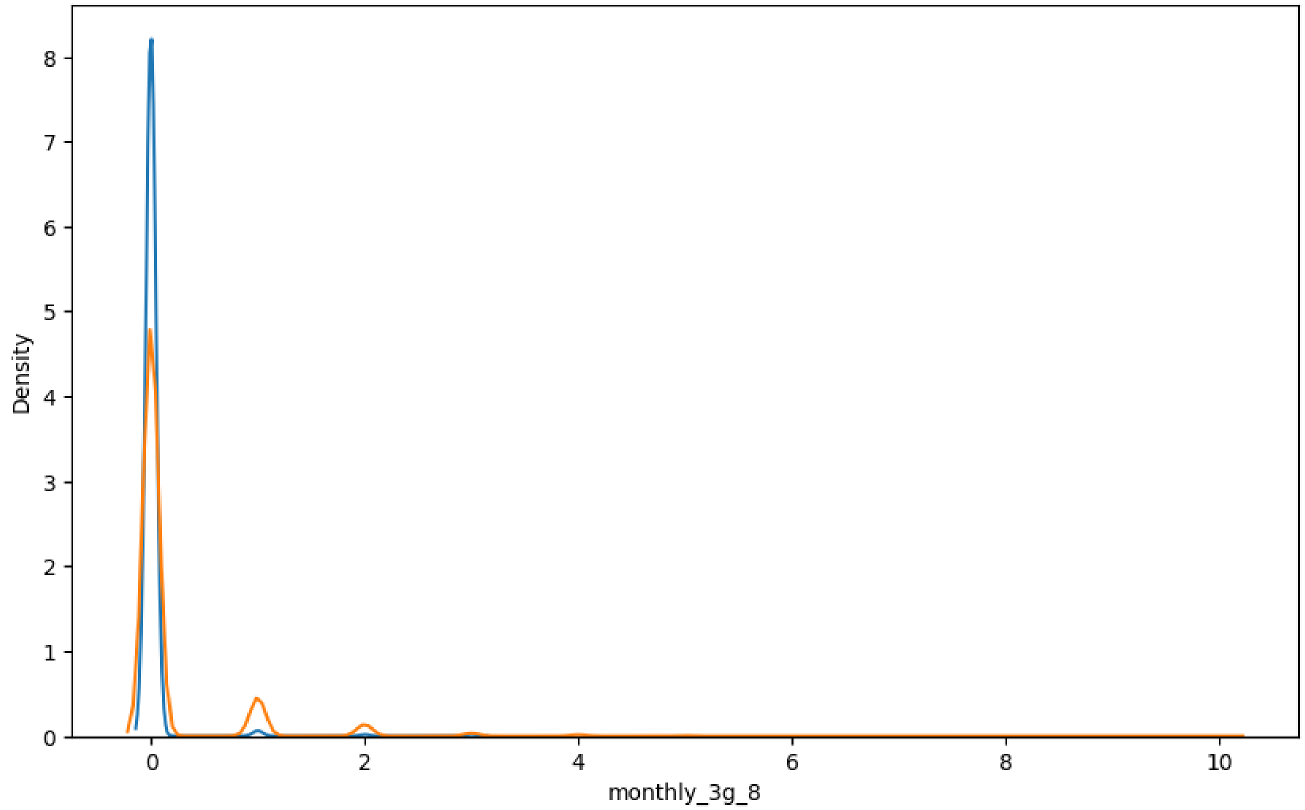
Important Predictor: ISD Outgoing Minutes of Usage in August



Insights:

It is evident that the ISD outgoing minutes of usage in August are nearly absent for churned customers, with values concentrated around zero. Conversely, for non-churn customers, these minutes of usage are slightly higher in comparison to churned customers.

Important Predictor: Monthly 3g data usage of August



Insights:

The number of monthly 3G data usage in August for churned customers is highly concentrated around the value of 1. In contrast, for non-churn customers, this variable exhibits a broader distribution across various values.

Business recommendations

- Target customers with reduced local incoming and outgoing ISD call usage in August.
- Focus on customers who show decreased outgoing charges to other operators in July and reduced incoming charges from other operators in August.
- Offer special incentives to customers experiencing an increase in value-based cost during the action phase.
- Monitor and engage with customers who reduce their monthly 3G usage in August.
- Identify and address customers with declining STD incoming minutes of usage for operators T to fixed lines of T in August.
- Pay attention to customers who decrease their monthly 2G usage in August.
- Engage with customers experiencing reduced incoming minutes of usage for operators T to fixed lines of T in August.
- For customers with increasing roaming outgoing minutes of usage (roam_og_mou_8), consider retention strategies to reduce churn.
- These recommendations can help the telecom company proactively retain high-risk customers and reduce churn.

End of Slides

Thanks for reviewing our Case Study on Telecom Churn.

With Regards,

Shahul Hameed – mashahulhameedh@gmail.com

Aishwarya Kumar Sharma – sharma.aishwarya093@gmail.com

Ashrit Gaikwad – ashritgaikwad5696@gmail.com