CS610 Applied Machine Learning

– a Prediction in Telecom Customer Churn

Group 4
Chen Fangxian
Chen Yiman
Tan Hui Ying, Amanda
Tay Zhi Xiang
Quek Zixin, Richard

Introduction

Business Problem

- Studies suggest that acquiring a new customer can cost 5-7 times more than retaining an old one
- To understand why the customer churn and what the important factors are

Dataset

- 1 month sample telecom churn data from IBM for customers based in different cities in California
  - > 7043 rows
  - > 33 columns
- Median income per capital by cities from US Census to retain some characteristics of data from different cities

Machine Learning Application

- Run models to determine which model gives the best binary prediction and its feature importance

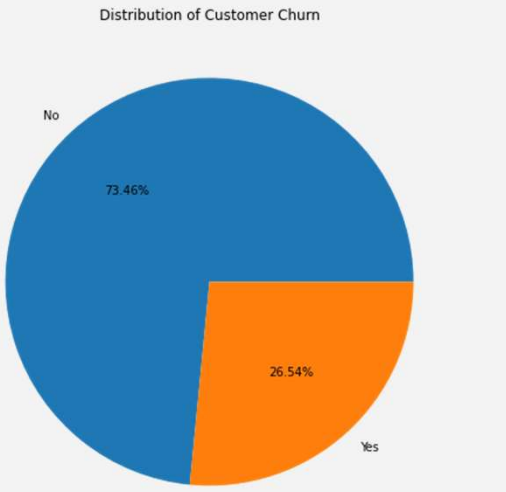
Data Preprocessing

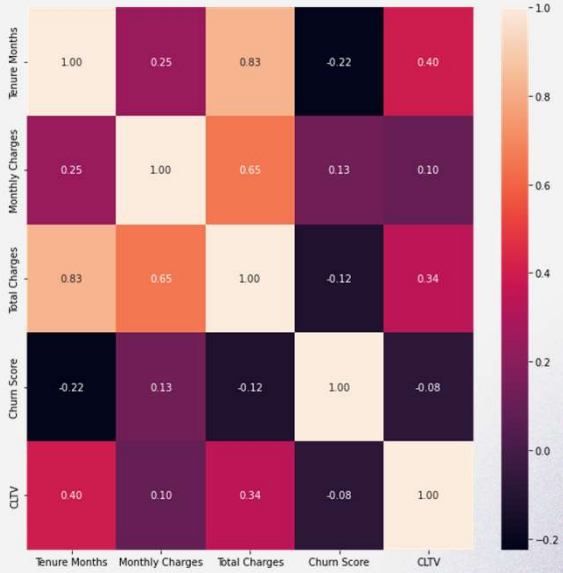
- **Imbalanced data**: the proportion of churn label “no” is 3x more than “yes”, this will impact the prediction, especially in our case, where we are doing binary prediction → Use SMOTE (Synthetic Minority Oversampling Technique) to generate new instances from existing data in a balance way

- **Irrelevant data**: Some of the churn reasons are not relevant such as “moved” and “deceased”, we removed the corresponding rows

- **Multicollinearity**: we discovered strong correlation between total charges with tenure months and monthly charges, we compare the logistics regression (LR) by dropping ‘Total Charges’

- **Imputation**: some missing data has been imputed with its median to not change the distribution



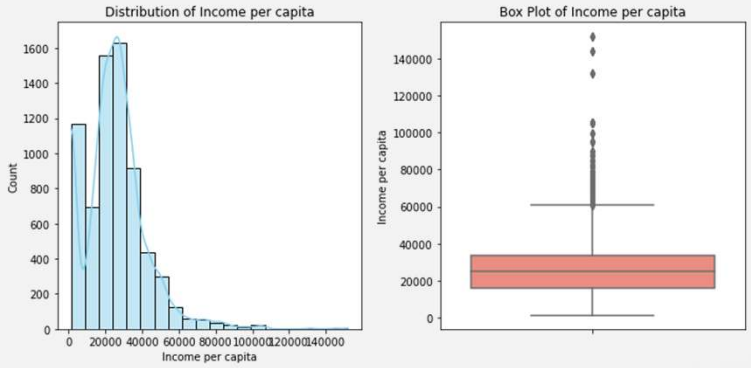


Data Preprocessing

- **One-hot encoding**: there are many columns are categorical such as gender, dependent, payment method, etc. We encoded them to numerical data to be able to run the models

- **Standardization**: after encoding, we realized that our model ran too perfectly, we realized it's due to columns such as Tenure, Charges and Income per Capita were out of the scales compared to the rest of the columns, hence, we standardized these columns to the similar scale.

- **Drop irrelevant columns** for model running: CustomerID, Count, Country, State, City, Zipcode, Lat Long, Latitude, Longitude, Churn Score, Churn Reason

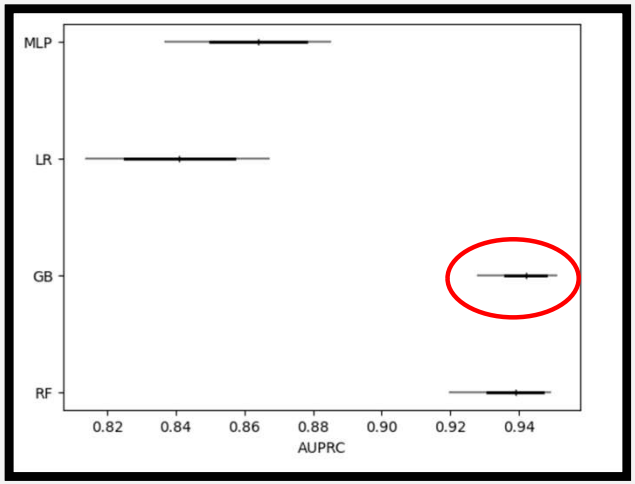


Used Features

Type	Features	Remarks
Discrete	• Gender, senior citizen, partner, dependents, tenure months, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming movies, contract, paperless billing, payment methods	• One-hot encoded based on the original values
Continuous	• Tenure, Monthly charges, Total Charges, CLTV, Income per capita	• The continuous features were standardized

ML Models Result Comparison

	Model	Threshold	AUROC	AUPRC	Sensitivity	Specificity	Runtime
0	RF	0.580	0.942 (0.928-0.949)	0.939 (0.92-0.949)	0.851 (0.846-0.911)	0.885 (0.823-0.897)	4.934
1	GB	0.543	0.936 (0.925-0.943)	0.942 (0.928-0.951)	0.85 (0.835-0.874)	0.862 (0.828-0.882)	2.219
2	LR	0.539	0.864 (0.85-0.877)	0.841 (0.814-0.867)	0.802 (0.77-0.831)	0.78 (0.755-0.826)	0.299
3	MLP	0.455	0.89 (0.875-0.899)	0.864 (0.837-0.885)	0.86 (0.819-0.869)	0.783 (0.765-0.811)	6.169

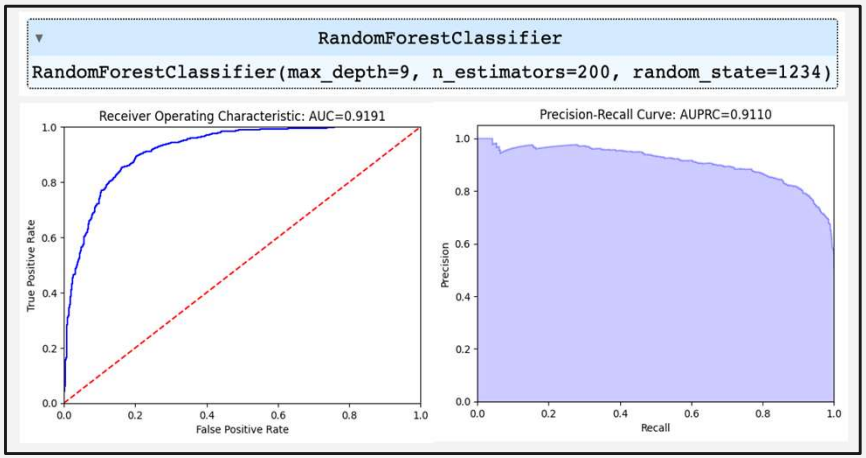


AUROC vs AUPRC?

- Balanced data & Bianry Prediction
- Focus on min.of False Positive Rate

- Finally choose Random Forest model for further hyperparameter tuning

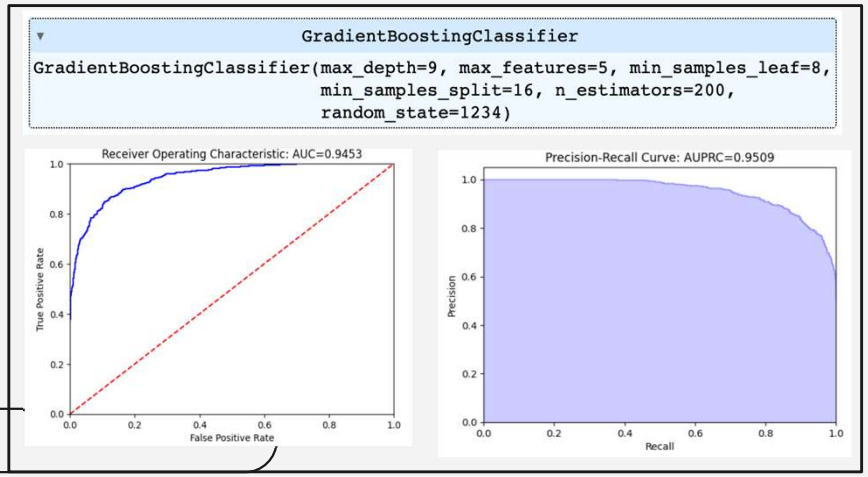
Hyperparameter Tuning



Random Forest:

Tuning Methods: GridSearchCV
Tuned: ‘max_depth’, 'min_samples_leaf’, ‘min_samples_split’, 'n_estimators’

ROC-AUC: 0.942 → **After Tuning** → **0.9191**

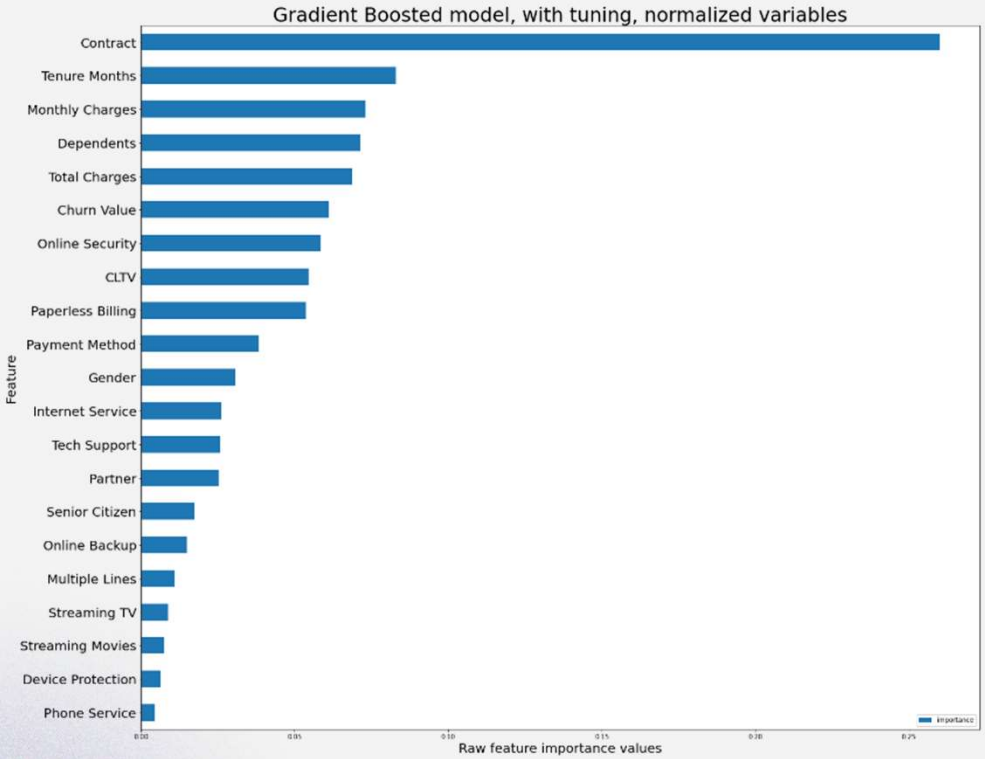


Gradiant Boosting:

Tuning Methods: GridSearchCV
Tuned: max_depth’, ’max_features’, 'min_samples_leaf’,min_samples_split’
'n_estimators’

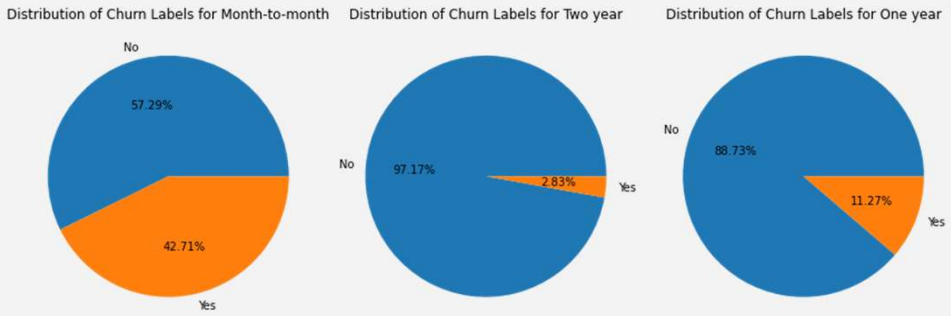
ROC-AUC: 0.936 → **After Tuning** → **0.9453**

Feature Importance



The below feature stands out more than others

- **Contract**: there is a likelihood that Month-to-month subscribers will easily churn as they has less contractual obligations if they decide to end



Conclusion, Limitations & Future Work

Conclusion:

- Random Forest model predicted the highest ROC, resulting in lowest False Positive Rate
- After tuning, ROC of RF decreased so we tried to tune GB which achieved highest ROC
- 4 most important features that contributed to churn are **Contract, Tenure Months, Monthly Charge, Dependent**

Limitation:

- Limited dataset: the data is only 1 month worth, with more data records, the model could be more robust
- Data could contain timestamp for a better churn prediction with time which we could identify the main factors that caused churn and to address them in a timely manner

Future Work:

- Based on feature importance, we can deep dive into more important features individually to find out the correlation of between the features and the churn, so we could allocate the right resources to retain the customers
- Other factors could be taken into consideration such as overall telecom landscape with competitor data (typically an organization could obtain from data agency), promotional mechanism to assist better understanding of churn vs acquisition cost

Q&A

Thank you!