Customer Churn Prediction on Telecom Data From Kaggle

Context:

- This particular churn prediction problem was hosted on Kaggle in 2020. You can find the original dataset and problem statement https://www.kaggle.com/c/customer-churn-prediction-2020/overview.
- Customer churn happens when a customer ceases to utilize the service provided by business.

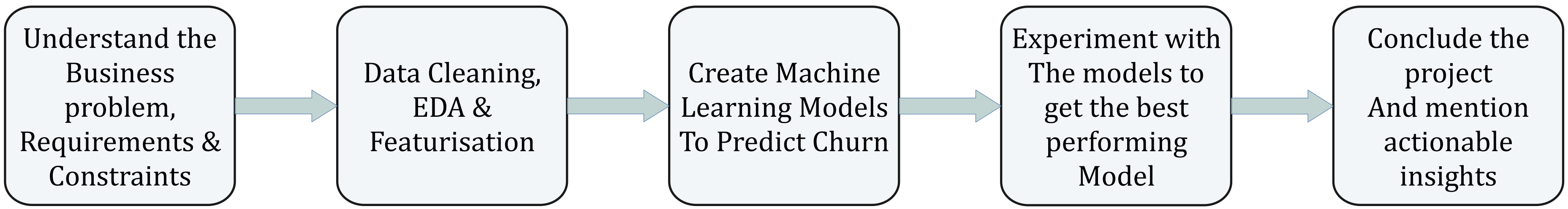
Business Motivation and Goal:

- Telecom industry has a huge problem of customer churn. Customers leave and join new service every now and then.
- The goal of this project is to analyse the churn data to get some actionable insights and help the business in preventing customer churn.

Metrics and Business Constraints:

- Their is no strict latency requirement
- High Accuracy is demanded
- Recall would be also a good metric to notice, along with the confusion matrix.

Methodology/Approach:



Available Features:

1.) state
2.) account_length
3.) area_code
4.) international_plan
5.) voice_mail_plan
6.) number_vmail_messages
7.) total_day_minutes
8.) total_day_calls
9.) total_day_charge
10.) total_eve_minutes
11.) total_eve_calls
12.) total_eve_charge
13.) total_night_minutes
14.) total_night_calls
15.) total_night_charge
16.) total_intl_minutes
17.) total_intl_calls
18.) total_intl_charge
19.) number_customer_service_calls
20.) churn

Exploratory Data Analysis And Featurisation

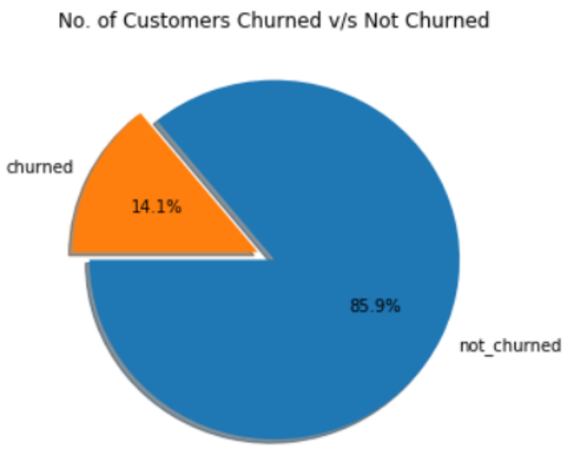


Fig. 1

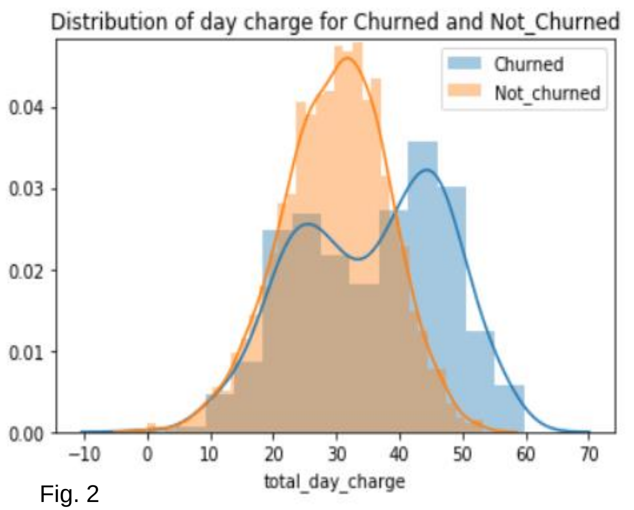


Fig. 2

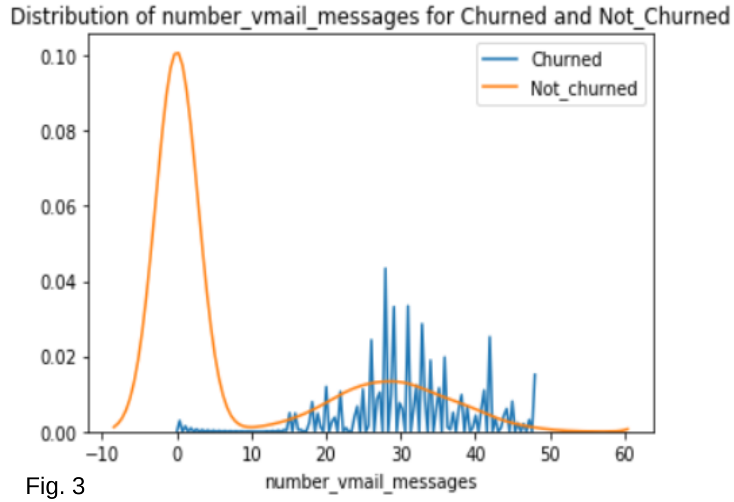


Fig. 3

EDA Highlights:

- Dataset is highly imbalanced (Fig. 1), number of churned examples is much lower than the not_churned examples.
- Features such as total_day_charge (Fig. 2), total_night_charge, total_eve_charge, total_intl_charge and number_vmail_messages (Fig. 3) are showing a decent separation between the churned and not_churned examples.
- Features representing minutes and charge are highly correlated.
- For more detailed EDA check the ipynb here: https://github.com/S-G-001/customer_churn_prediction

Featurisation:

- One hot encoding was done for categorical features.
- Features have been standardised by removing the mean and scaling the variance.
- Features that were Highly correlated with other features have been removed from the training process. Removed features are: 'total_day_minutes','total_eve_minutes','total_night_minutes','total_intl_minutes'.

Machine Learning Models And Their Results

Models:

•Five classification models were trained - two SVM, one Random Forest and two XGBoost classifiers. Below tables lists their performance on training data and test data.

S.No.	Model	class imbalance status	Train_accuracy	Test_Accuracy	Test_Recall_score
1.	SVM	No class balancing	0.98	0.91	0.48
2.	SVM	Balanced using class weights	0.98	0.9	0.65
3.	RF	Balanced using class weights	0.97	0.92	0.72
4.	XGBClassifier	Balanced using scale_pos_weight	1.0	0.95	0.75
5.	XGBClassifier	Balanced using SMOTE	0.99	0.84	0.87

Conclusion:

•XGBClassifier (Model 4) is giving the highest test accuracy and a decent recall score.

•Test recall score is highest for XGBClassifier(Model 5) when data is balanced using SMOTE, but it's leading to poor test accuracy.

•4th model with 95% accuracy and 75% recall seems to be the best choice for predicting churn.

Key Action Points:

•Some of the most important features in prediction of Churn are: Total_day_charge, number_customer_service_calls, total_eve_charge, total_night_charge and international_plan. Marketing/business team can act on these features to prevent churn.

•Main challenge of this project is the class imbalance, to get a better recall score we need more data points belonging to minority class. And more data in general will enable us to build deep learning models and get better test accuracy.