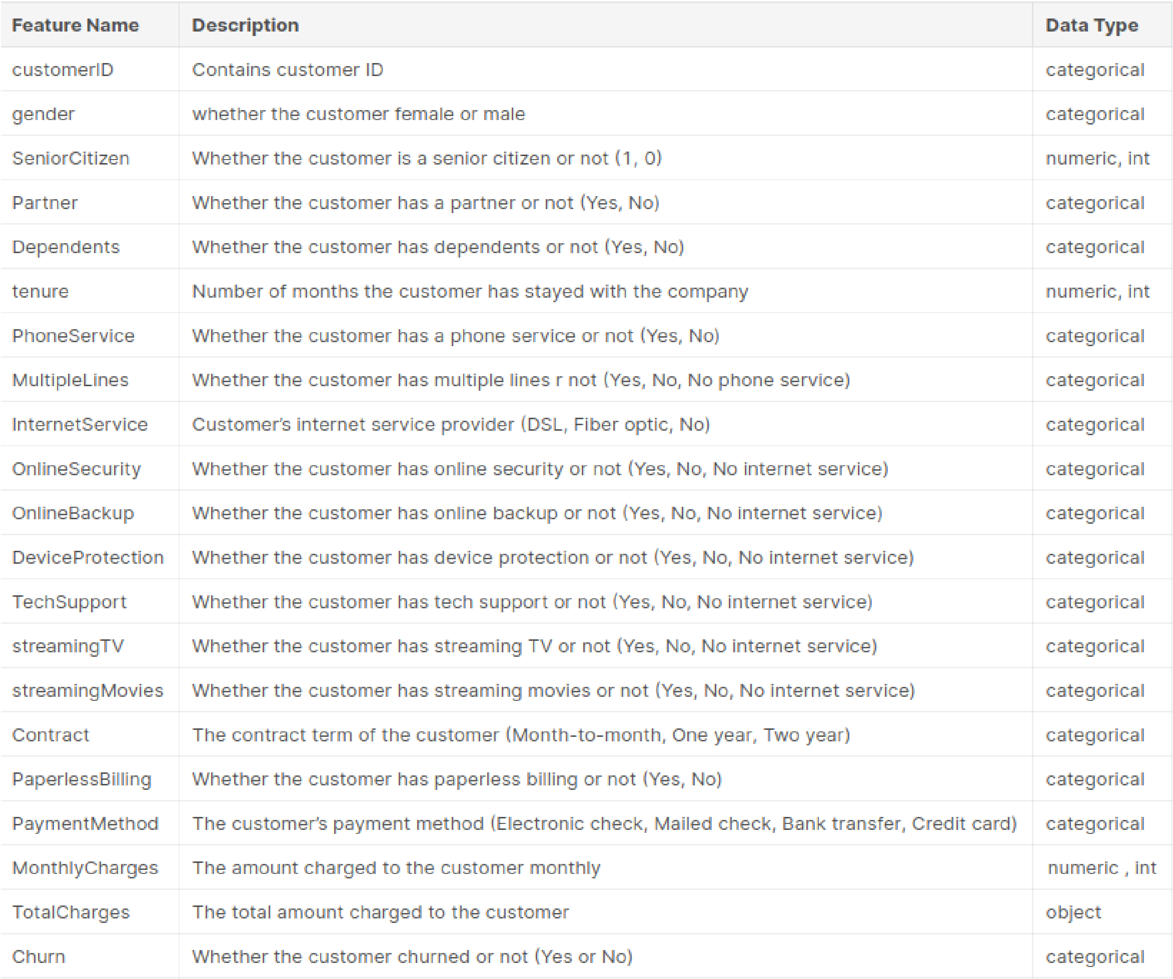
TELECOM CUSTOMER CHURN PREDICTION

1. INTRODUCTION

Customer churn is a major problem and one of the biggest concerns for telecommunication industry. Due to the cost of retaining an existing customer is much lower than acquiring a new one, the company is seeking to develop a churn prediction model to assist telecom operators in predicting which customers are most likely to lose. Therefore, identifying the factors that increase customer churn is important to build this model. The purpose of this project is to explore these data more deeply, utilizing nonparametric statistical methods to do so. Through the course of this analysis, new insights will be offered as to the types of indicators that influence churn and charge, as well as attempting to compare both nonparametric and parametric methods.

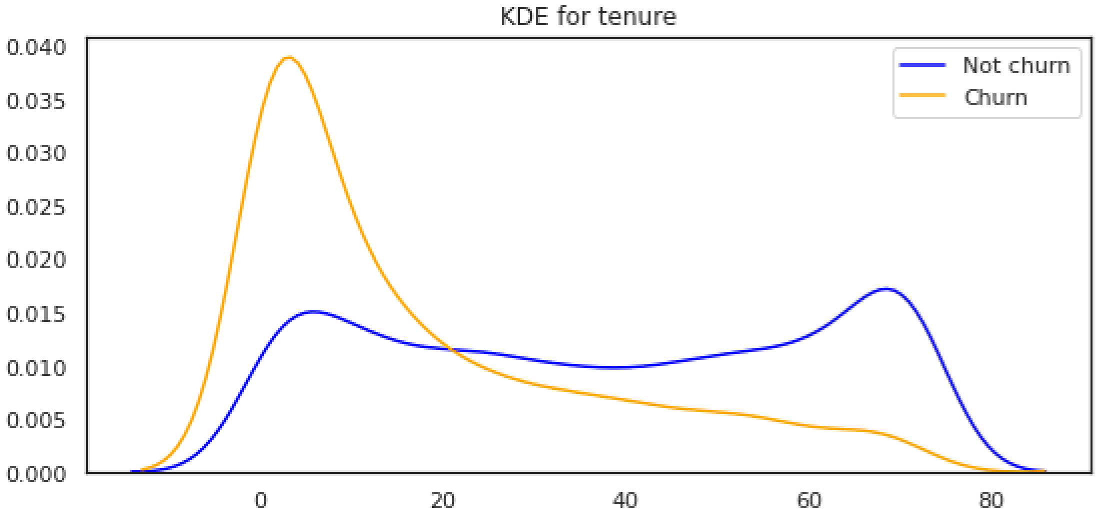
2. DATA EXPLORATION

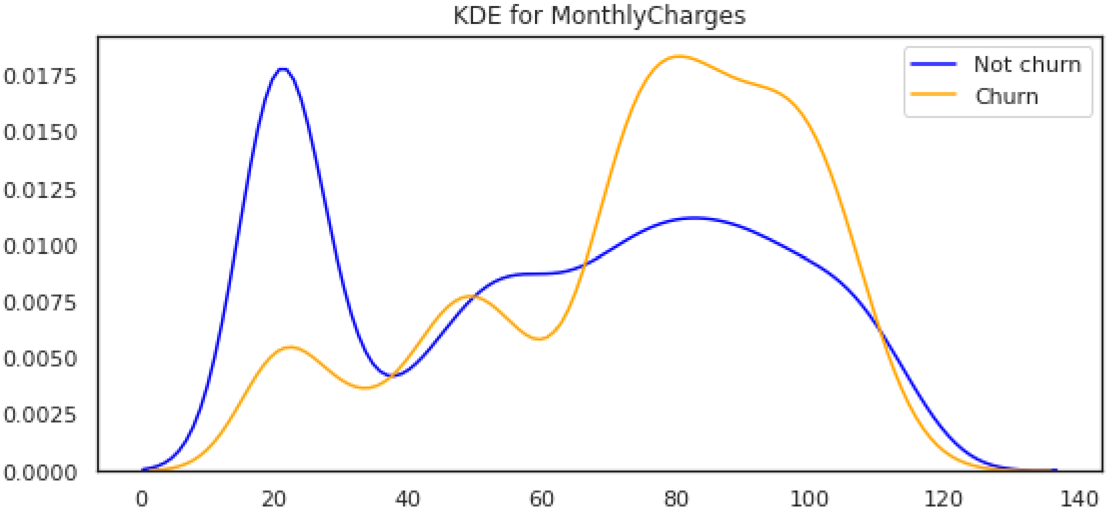
The raw data contains a total of 7043 observations and 21 variables. Description and data types of each features in dataset;

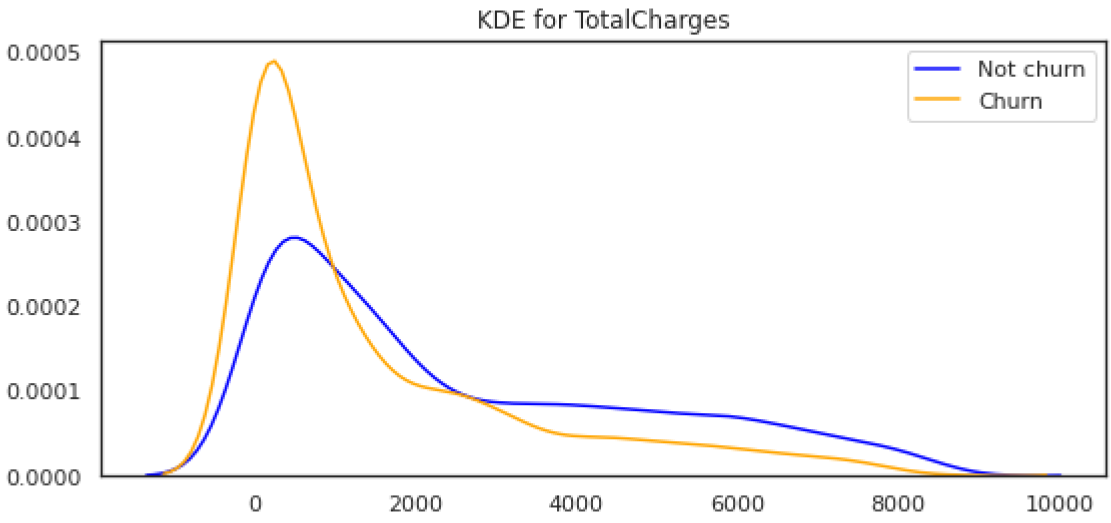


Feature Name	Description	Data Type
customerID	Contains customer ID	categorical
gender	whether the customer female or male	categorical
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)	numeric, int
Partner	Whether the customer has a partner or not (Yes, No)	categorical
Dependents	Whether the customer has dependents or not (Yes, No)	categorical
tenure	Number of months the customer has stayed with the company	numeric, int
PhoneService	Whether the customer has a phone service or not (Yes, No)	categorical
MultipleLines	Whether the customer has multiple lines r not (Yes, No, No phone service)	categorical
InternetService	Customer's internet service provider (DSL, Fiber optic, No)	categorical
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	categorical
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	categorical
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)	categorical
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)	categorical
streamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)	categorical
streamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)	categorical
Contract	The contract term of the customer (Month-to-month, One year, Two year)	categorical
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)	categorical
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer, Credit card)	categorical
MonthlyCharges	The amount charged to the customer monthly	numeric , int
TotalCharges	The total amount charged to the customer	object
Churn	Whether the customer churned or not (Yes or No)	categorical

However, not all of the predictors are useful for our models, specifically customerID, and thus we remove it first. We also have some missing values in some features. There are a total of 11 missing values in TotalCharges column. All of the missing values have churn=”No” feature, which is over-represented in the data. We will just delete those rows from the working data. The dataset contains tenure, MonthlyCharges and TotalCharges numerical features. We will mostly use this features on data exploration. Probability density distributions and statistics about this features;

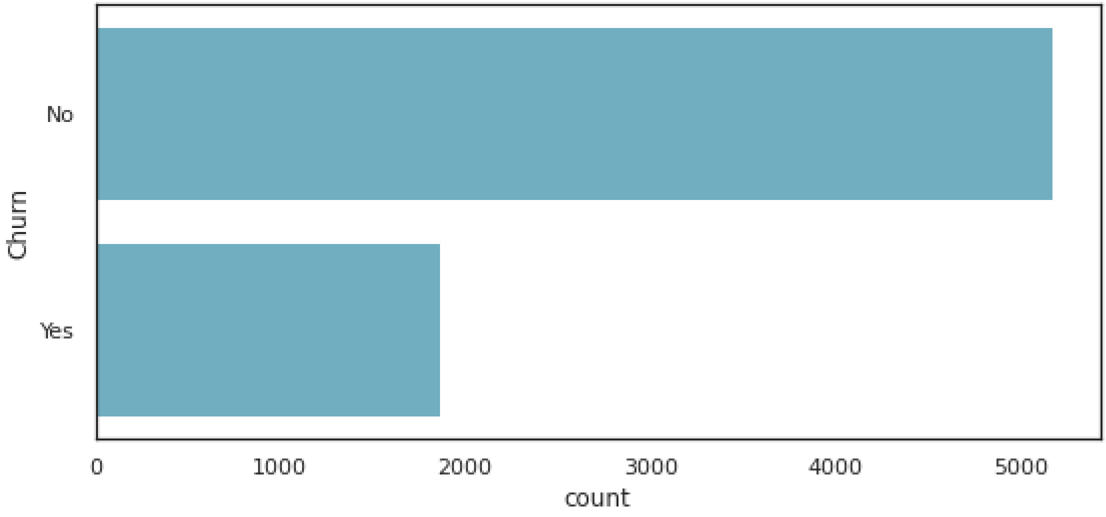




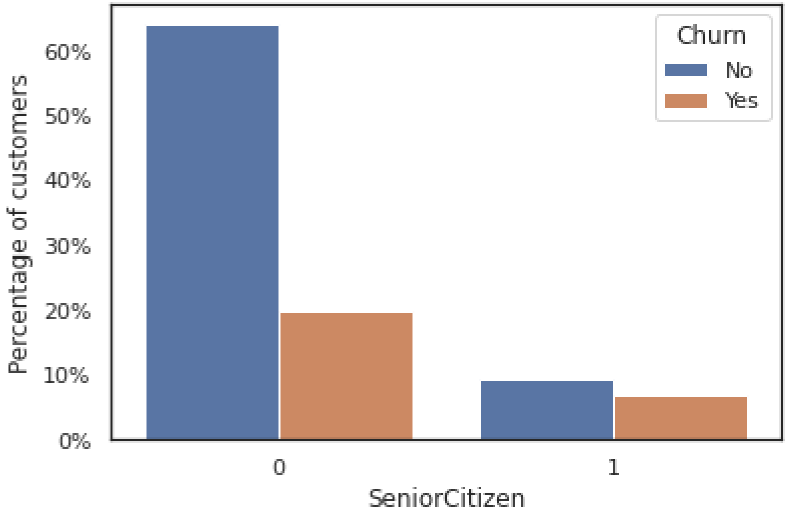


From the plots above we can conclude that, recent clients are more likely to churn. Clients with higher MonthlyCharges are also more likely to churn, tenure and MonthlyCharges are probably important features. We can create categorical column for tenure feature to show customer attrition by tenure groups.

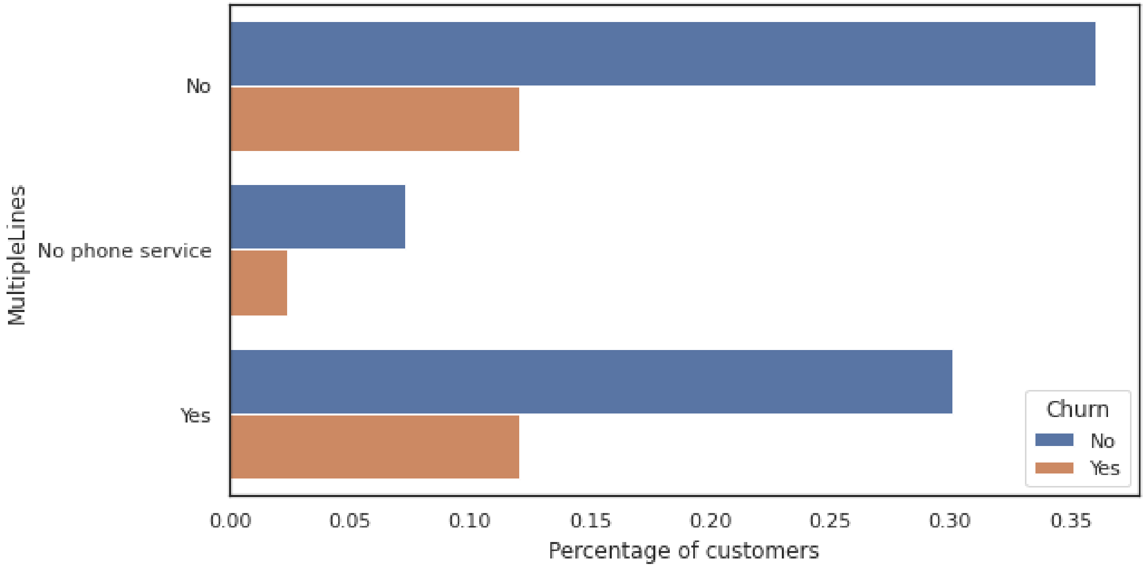
We are trying to predict if the client left the company in the previous month. This is a binary classification problem with a slightly unbalanced target variable. Churn customer and non-churn customer numbers;

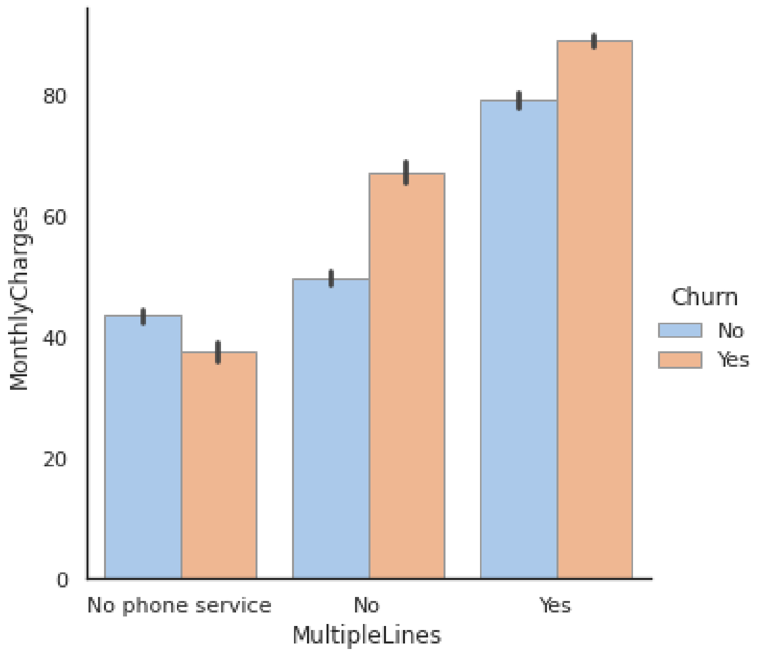


There are 16 categorical features: 6 binary features (Yes/No), 9 features with three unique values, 1 feature with four unique values. Churn percentage of customers by gender and age (SeniorCitizen feature) are represented in this graph;

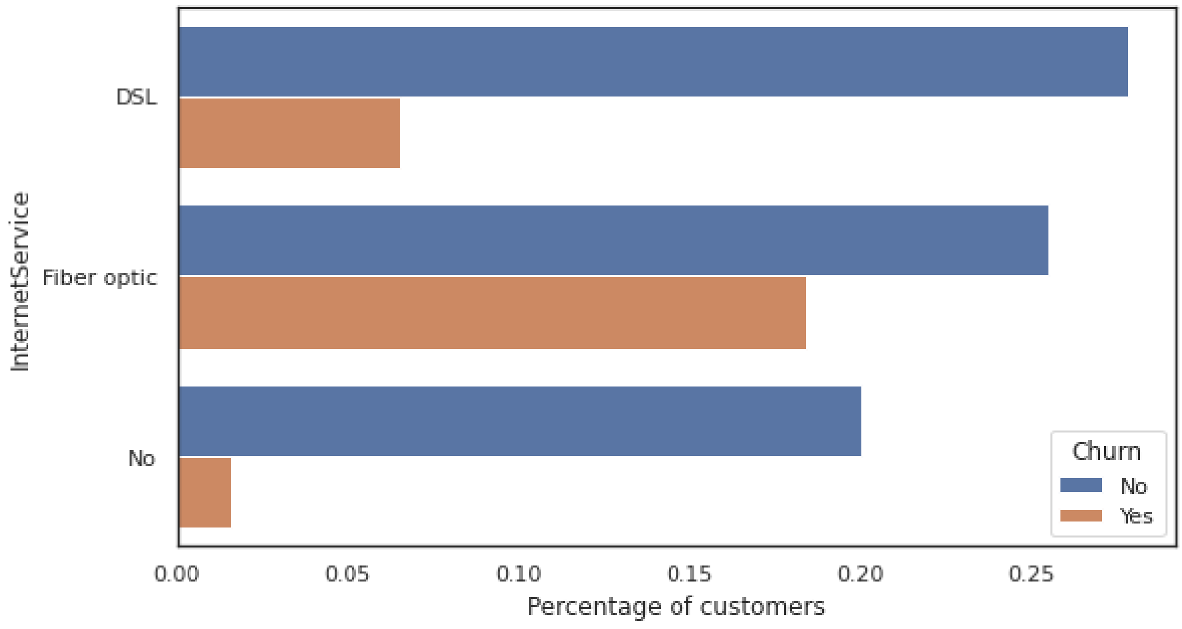


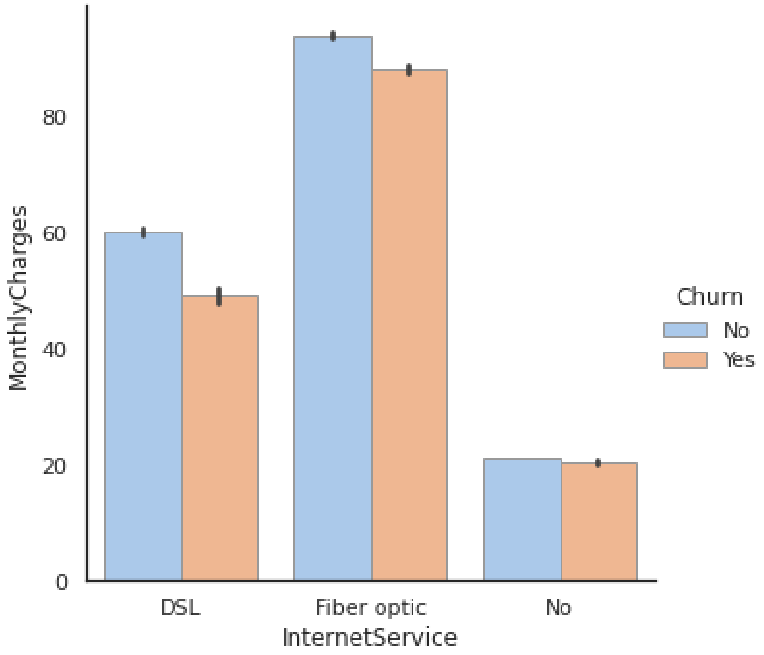
From the plots above that we can conclude that, gender is not an indicative of churn. Senior citizens are only 16% of customers, but they have a much higher churn rate: 42% against 23% for non-senior customers. There are no special relations between this categorical values and the main numerical features. Churn percentage of the clients that have phone services and have more than one lines visualized in this charts;



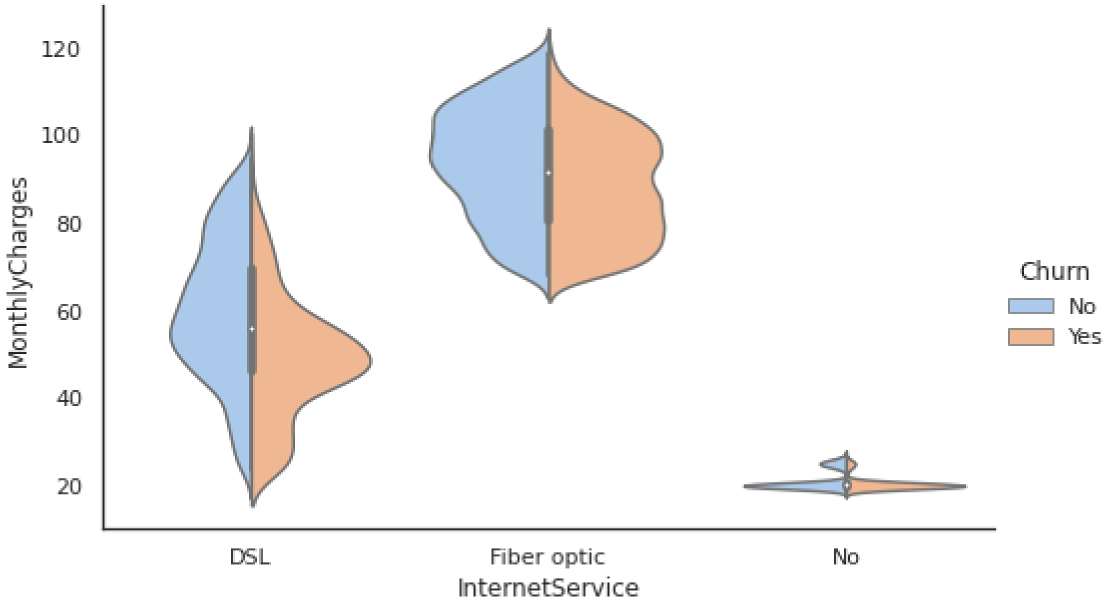


From the plot above that we can conclude that, just a few customers does not have phone services. Customers with multiple lines have a slightly higher churn rate. Effects of monthly charges and Internet services features;





The clients without Internet services have a very low churn rate. Customers with fiber optic service are more probable to churn than those with DSL connection. Comparing the Internet service with monthly charges;



There are various plots in graphs folder for other categorical features.

3. FEATURE SCALING AND MODEL BUILDING

Before model building, all variables scaled a range of 0 to 1 and transformed. Data splitted to training and test sets. Test sizes are 0.3 or 0.25 for different classification methods. 11 different models builded using logistic regression classification, AdaBoost classifier, XGBoost classifier, Gaussian Naive-Bayes classifier, decision tree, k-nearest neighbor classifier, multi-layer perceptron neural network classifier, neural networks with Keras, random forest classifier, support vector machine and gradient boosting classifier.

4. PERFORMANCE OF MODELS

Performance of models are tested with confusion matrix, accuracy score, ROC-AUC score, ROC curve, f1-score, precision score, recall score and Cohen’s Kappa coefficient metrics. Best performance are obtained with artificial neural network classifier using Keras and TensorFlow. All results are plotted in graphs folder.

5. CONCLUSION

From exploratory data analysis we can conclude that, recent clients are more likely to churn and the clients with higher MonthlyCharges are also more likely to churn. Senior citizens are only 16% of customers, but they have a much higher churn rate: 42% against 23% for non-senior customers. There are no special relations between this categorical values and the main numerical features.

Additional services variables (OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies) are plotted with churn rates and these results obtained; customers with the first 4 additional services (OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport) are more unlikely to churn and streaming service is not predictive for churn. Payment methods plotted with churn rates and these results obtained; customers with paperless billing are more probable to churn and the preferred payment method is electronic check with around 35% of customers. This method also has a very high churn rates. Short term contracts have higher churn rates. Longer contracts are more affected by higher monthly charges. Mailed checks have lower charges There is a huge gap in charges between customers that churn and those that do not with respect to mailed check.