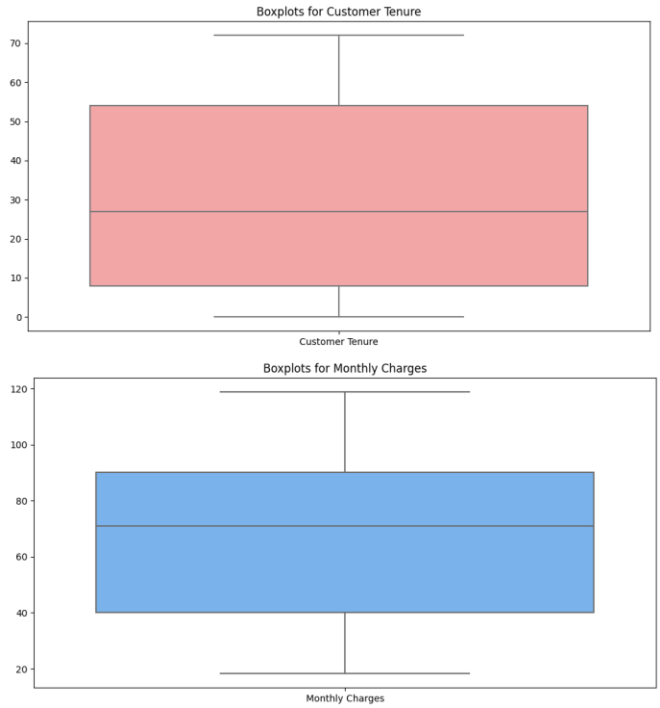
BA2 – Assignment 1 – Report – Sami Seppälä

Business Understanding

Customer churn is the portion of customers that cancel their subscription to our service. As acquiring new customers is more expensive than keeping existing customers, it is essential to reduce churn as much as possible within our business. The more paying customers we have the more profitable our business will be. If we are able to increase the tenure of customers our business becomes more stable as the longevity of customers and therefore our cash flow increases. If we were able to predict customer churn and identify features that affect it, we could take targeted actions in order to decrease the churn rate within e.g. specific segments. By identifying the underlying aspects affecting churn we would be able to create specific targeted marketing campaigns or targeted customer service in order to combat churn within specific segments and thus increase our revenue.

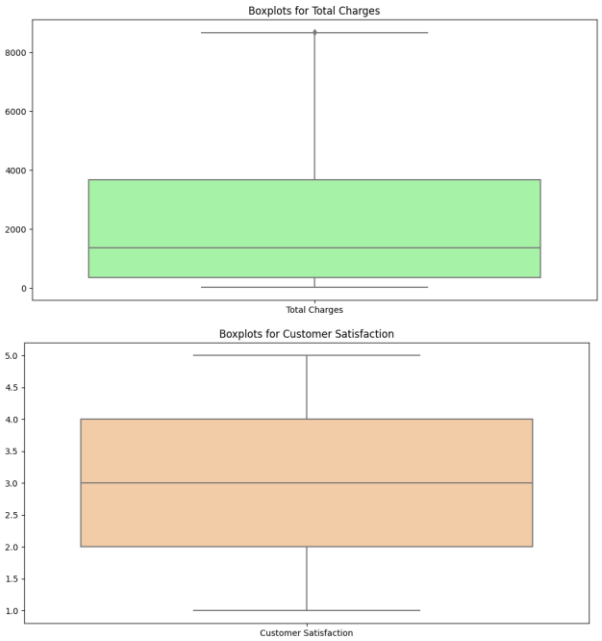
Data Understanding

In order to predict we have different forms of data at our disposal in addition to the churn numbers themselves. All the data on individual customers is gathered in the churn_dataset.csv file which will be used for the analysis. We have information on customer gender, age, partner status, dependents, tenure, subscription info, streaming info, and payment information. Before any analysis is done some features that obviously will affect customer churn will be variables like their contract length and tenure as a customer. Some missing values within 'Total Charges' are present for customers with a tenure of less than one month. These values will be replaced by the monthly charges. I created some boxplots in order to analyse the distribution of the numerical values:

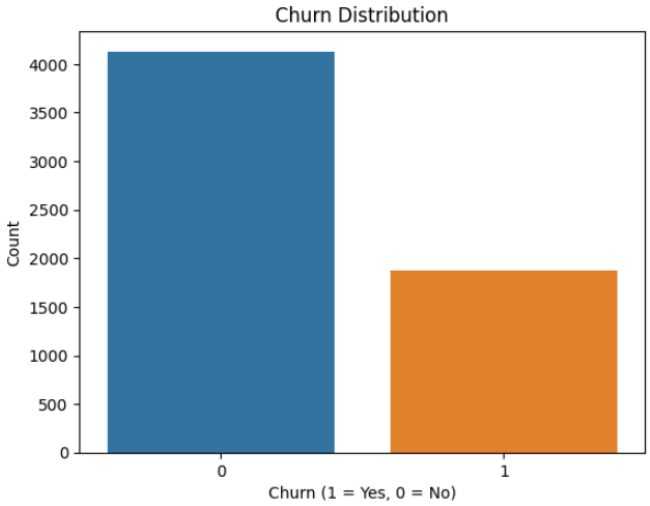


Further analysis using IQR boundaries of 1.5 for the 25th and 75th percentile of data only showed one outlier row which was

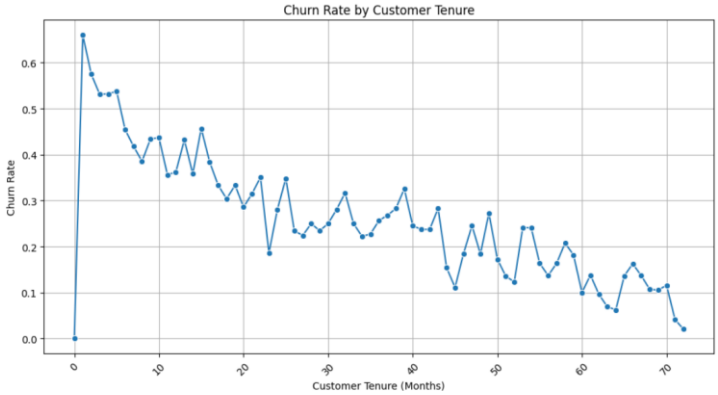
thus decided to be kept in the data. The Interquartile Range is the difference between the first and third quartile within a dataset.



One-hot encoding was perfumed on categorical features in order to compare them statistically with the target variable. Analysing the relationship between features and the target variable 'Churn' showed that the variables with highest positive correlation were a fibre optic internet plan, electronic check as payment method and paperless billing. The features with the highest negative correlation were customer tenure, two year contracts and customers with dependents. The churn rate also showed that there were many more customers who did not churn (~4000) compared to churned customers (~2000).



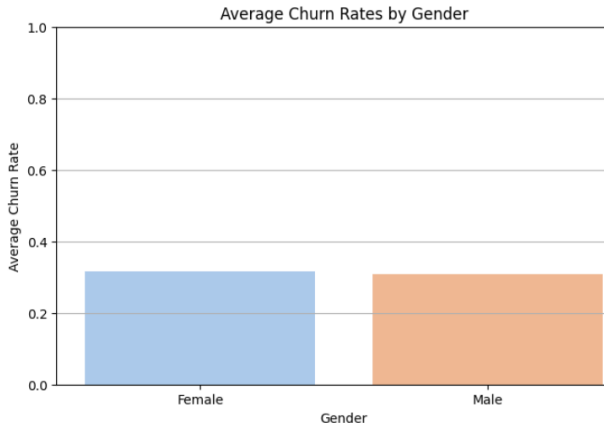
The churn rate also seems to decrease as customer tenure becomes longer which proves the effect of trying to hold on to customers for as long as possible.

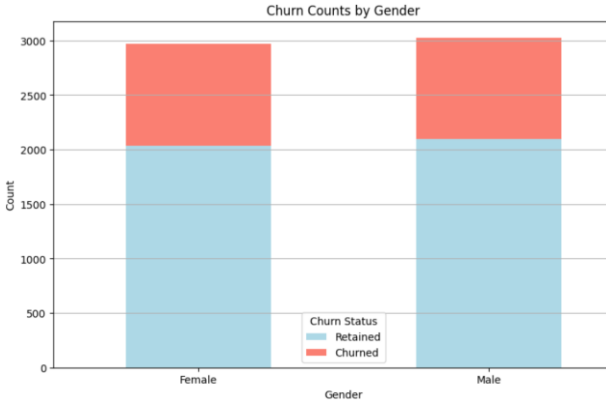


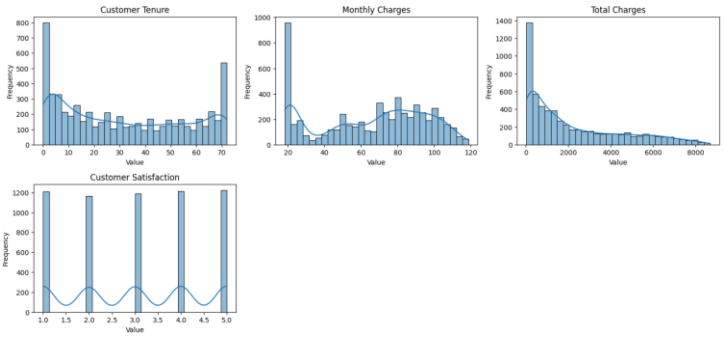
The individual churn rates for the encoded variables did not show any distinct differences between them.

	Feature	Average Churn Rate
0	Customer Tenure	0.258674
1	Monthly Charges	0.313323
2	Total Charges	0.298316
3	Customer Satisfaction	0.311523
4	Customer Gender_Male	0.311541
5	Is Senior_Y	0.376672
6	Partner_Y	0.307845
7	Dependents_Y	0.228017
8	Has Phone Service_Y	0.308430
9	Multiple Lines_No phone service	0.308430
10	Multiple Lines_Y	0.314061
11	Internet Plan_Fiber optic	0.324528
12	Internet Plan_No	0.230025
13	Online Security_No internet service	0.230025
14	Online Security_Y	0.269944
15	Device Protection_No internet service	0.230025
16	Device Protection_Y	0.300389
17	Tech Support_No internet service	0.230025
18	Tech Support_Y	0.272444
19	Streaming TV_No internet service	0.230025
20	Streaming TV_Y	0.318375
21	Streaming Movies_No internet service	0.230025
22	Streaming Movies_Y	0.318015
23	Contract_One year	0.246420
24	Contract_Two year	0.213700
25	Paperless Billing_Y	0.292061
26	Payment Method_Credit card (automatic)	0.264313
27	Payment Method_Electronic check	0.358266
28	Payment Method_Mailed check	0.282266

Gender did not seem to have an adverse affect on the churn rate as demonstrated by the histograms below.







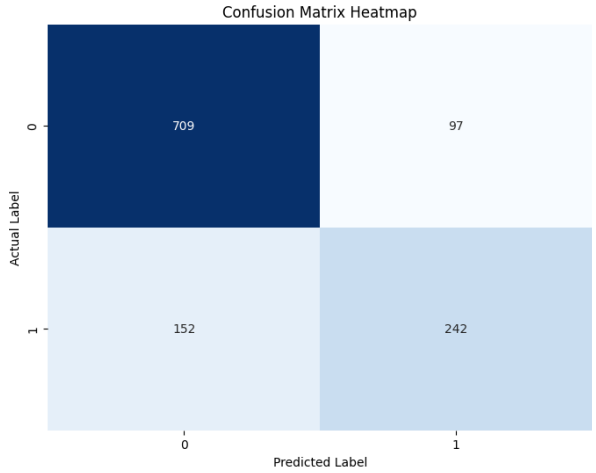
As the numerical features do not follow a gaussian distribution we will not perform standardization. As no outliers were removed normalization will also not be done as it is sensitive to discrepancies for outliers. In order to try to improve the accuracy of the model I will remove features with a close correlation to 0 with a cutoff value of 0.2 and -0.2 in order to only preserve relevancy. No features had a high correlation with the target variable so there is no fear of redundancy.

Model Building/Evaluation

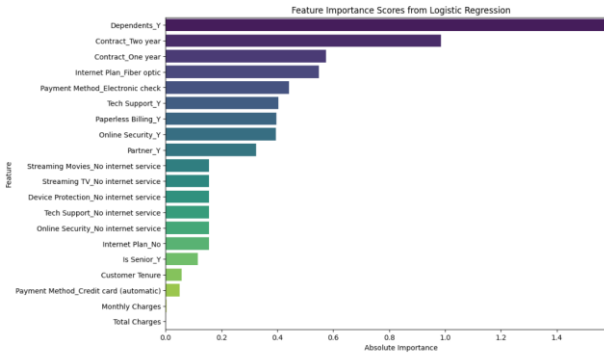
The cleaned up data is split into a training set (20%) and a test set (80%). This is done in order to test how well the trained models are able to generalize to unseen data. As there is a large difference in the number of retained and churned customers we will not use the accuracy metric which can suffer from unbalanced datasets. Instead the ROC AUC Score (Receiver Operating Characteristic Area Under the Curve) will be used. The ROC AUC Score is a metric that shows how well the model can differentiate between positive and negative

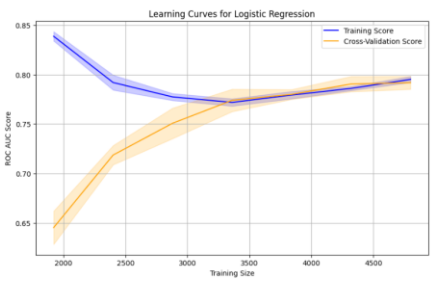
classes across all thresholds. A ROC AUC Score of 1 would mean the model would always be able to differentiate between churned and retained customers given a random customer while a ROC AUC Sore of 0.5 would be equal to random guessing. A baseline logistic regression model is trained to compare the other models to.

```
Logistic Regression Model:
Confusion Matrix:
[[709  97]
 [152 242]]
```




```
Logistic Regression Model:
Accuracy: 0.7925
ROC AUC Score: 0.8582443224043027
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.88      0.85       806
           1       0.71      0.61      0.66       394

    accuracy                           0.79      1200
   macro avg       0.77      0.75      0.76      1200
weighted avg       0.79      0.79      0.79      1200
```





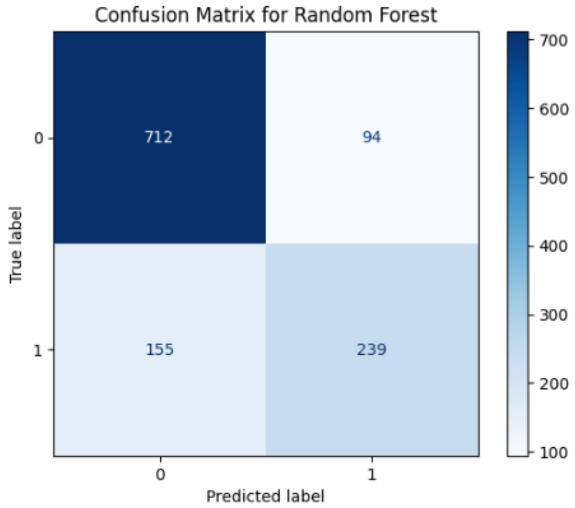
The logistic regression model achieved a ROC AUC Score of 0.858. The training curve shows that the model achieves a high accuracy for its cross-validation score converging with the training score proving that no over fitting is present. Next three new models are trained and tested using Decision Tree, Bagging with decision tree and Random Forest methods. Strengths of the decision tree model include how simple it is to understand while having no need for scaling of features. As a simple model it could be prone to overfitting, however. Bagging using multiple decision trees helps to reduce variance and improve accuracy by increasing the number of decision trees used and giving an overall accuracy rating. By being more complex the model as a whole could thus be harder to understand. The Random Forest method provides information on the predictive importance of different features while simultaneously reducing overfitting by using ensemble learning in combining multiple decision trees.

```
Decision Tree model:
ROC AUC Score: 0.6710348150294114
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.80      0.79       806
           1       0.57      0.54      0.55       394

    accuracy                           0.71      1200
   macro avg       0.67      0.67      0.67      1200
weighted avg       0.71      0.71      0.71      1200
```

```
Bagging Model:
ROC AUC Score: 0.8352127445176405
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.87      0.84       806
           1       0.68      0.57      0.62       394

    accuracy                           0.77      1200
   macro avg       0.74      0.72      0.73      1200
weighted avg       0.76      0.77      0.76      1200

Random Forest model:
ROC AUC Score: 0.8407958710685091
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84       806
           1       0.69      0.56      0.62       394

    accuracy                           0.77      1200
   macro avg       0.75      0.72      0.73      1200
weighted avg       0.77      0.77      0.77      1200
```

The highest ROC AUC Score was achieved using the Random Forest Model. Next, we will implement parameter optimization using grid search in order to try to improve the accuracy of the model. The following results present the Random Forest model with the highest ROC AUC Score:

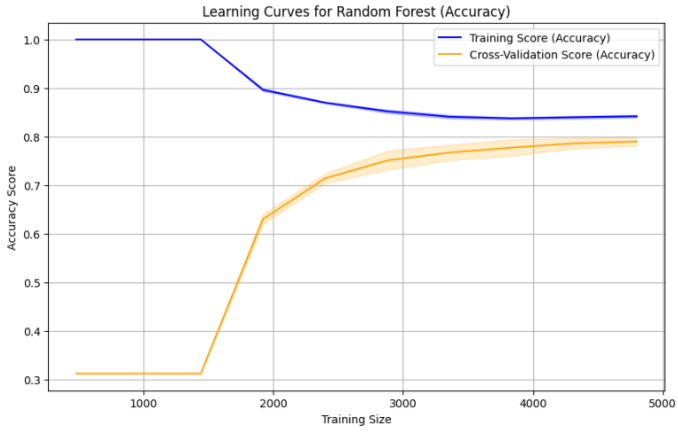
```
Optimized Random Forest Model:
ROC AUC Score: 0.8642997946870551
Confusion Matrix:
[[712  94]
 [155 239]]
```



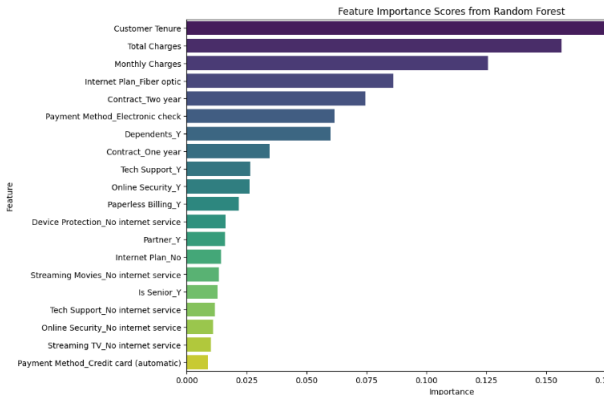

```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.88      0.85       806
           1       0.72      0.61      0.66       394

    accuracy                           0.79      1200
   macro avg       0.77      0.74      0.75      1200
weighted avg       0.79      0.79      0.79      1200
```

The model with the highest ROC AUC Score of 0.864 was a Random Forest model with the following parameters: {'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': 10, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 10, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 200, 'n_jobs': None, 'oob_score': False, 'random_state': 100, 'verbose': 0, 'warm_start': False}

As a parameter optimized Random Forest model the model performed the best since using ensemble learning methods reduce overfitting and increase the generalization capabilities of the model. By randomizing feature selection within the different decision trees Random Forest can reduce the effect of noise in the data. The learning curve shows converging training and cross-validation scores which show the lack of overfitting and support the ability of the model to generalize well to unseen data.



The four most important predictors in the final Random Forest model were customer tenure, total charges, monthly charges and a fibre optic internet plan.



Deployment

In order to implement the model into our business environment we first need to take the following steps: Plan the integration with existing systems and ensure a steady flow of data, actually deploying the model and all its corollaries, and monitoring the model and how it performs and making necessary adjustments to it. In order to implement the model with our Customer Management System we need to first ensure the availability and flexibility of our data pipeline. If the development of a user interface is not done in house, we might need to purchase new software that will allow the easy integration of machine learning applications into our existing systems. When the system is setup and running, we will be able to involve our marketing and customer service departments in order to start to combat churning. By identifying customers who are likely to churn targeted efforts can be made in order to get them to continue their subscription. This could include offering discounts or other incentives. Continuous monitoring of the model and the data it receives will be need in order to ensure no overfitting is happening. Data drift where customer behaviour changes over time also needs to be taken into consideration and the model adjusted accordingly. To ensure the model has not been biased in its training and ensuring fairness in treatment of customers across all segments it is important to evaluate the model for any potential bias or discrimination of certain customer groups.