DATA AND NETWORK MINING

Module Code: B8IT156

Names: Jack Maguire; Sean Carroll; Shane Crowley

Student Numbers: 20026112; 20024157; 20024425

Group: 2

Business Problem Framing

The chosen dataset for this project was the “Customer Churn” (Kumar, 2020) dataset sourced from Kaggle.com. This synthetic dataset is based on a fictitious telecom company. As per Ranjan et al. (2023) “One of the most important tasks that the telecom industry has to deal with is predicting churn.” The dataset is a collection of data in relation to customer level information for a telecom company, with customer use of ten different services recorded.

The primary business problem is to create a model using this dataset that accurately predicts customer churn within a telecom company. The model should aim to have a high sensitivity score to maximise the prediction of customers who churn as sensitivity “measures the proportion of positives that are correctly identified as such” (Dublin Business School, 2024). With the aim of telecom companies to retain as many customers as possible there is a need to analyse this dataset to explore how accurately customer churn can be predicted. This will aid stakeholders in making more beneficial decisions for their company. The stakeholders in relation to this business problem include company management, the marketing team, the sales team, investors, and data analysts.

An analysis of the dataset can provide valuable insights for telecom companies, making it a suitable dataset for an analytics solution. These insights contribute to business benefits, helping the company to achieve business objectives. As per Baremetrics (n.d.) the business could benefit from predicting customer churn because of the following:

- Accurate churn prediction aids in understanding future revenue by foreseeing potential customer losses.
- Forecasting individual churn rates enables targeted retention efforts, crucial for minimising customer attrition.
- Given the high cost of acquiring new customers, prioritising retention efforts is financially beneficial.
- Churn prediction facilitates identification of customer service shortcomings, allowing for targeted improvements.

The business problem is informed by the constraints in relation to this project, such as data quality, imbalance, regulatory compliance, and ethical implications that ensure the model's effectiveness in real-world applications. The primary objective of this project is to develop a predictive model for customer churn while taking these constraints into consideration.

Analytics Problem Framing

The analytics problem includes designing a machine learning model that accurately predicts customers who churn in a telecom company. Therefore, the desired research question to be answered is “How accurately can customer churn within a telecom company be predicted?” The designed model must navigate constraints such as data quality, imbalance, regulatory compliance, and ethical implications.

The Customer Churn dataset consists of over 3,334 instances and contains 11 features. The ABT contains the following attributes and outputs:

- **Churn:** The target variable. Does the customer churn or not (Categorical).
- **AccountWeeks:** The number of weeks the customer has had an active account (Discrete).
- **ContractRenewal:** If the customer recently renewed their contract or not (Categorical).
- **DataPlan:** If the customer has a data plan or not (Categorical).
- **DataUsage:** Gigabytes of monthly data usage (Continuous).
- **CustServCalls:** The number of calls into customer service (Discrete).
- **DayMins:** The average daytime minutes per month (Continuous).
- **DayCalls:** The average number of daytime calls (Discrete).
- **MonthlyCharge:** The average monthly bill (Continuous).
- **OverageFee:** The largest overage fee in the last 12 months (Continuous).
- **RoamMins:** The average number of roaming minutes (Continuous).

In addressing the business problem of predicting customer churn, it is assumed that the provided dataset accurately represents customer behaviour, with relevant and consistent features. It is also assumed that underlying churn factors remain stable over time, observations are independent, and business processes remain consistent. Additionally, it is assumed that the developed model can be applied to new customers effectively, and interventions based on churn predictions will mitigate churn rates efficiently.

The definitions of the key metrics of success, based on the measures of progress by Larson (2022) include:

Delivering Business Value: Translating customer churn data into business value.

Customer Centricity: Giving insights into the services that can be improved for customer satisfaction.

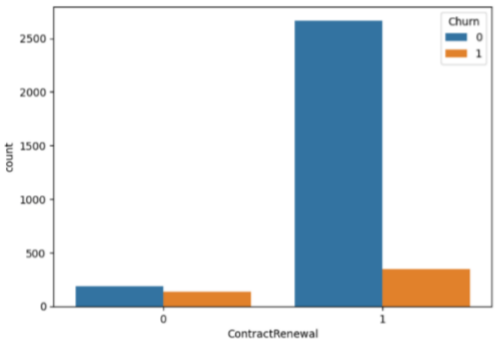
Analytic Operations: Building an explainable, ethical, and sustainable model that predicts customer churn.

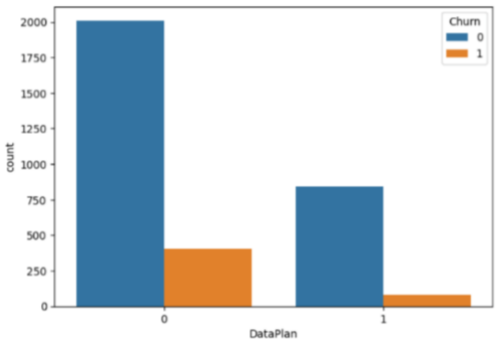
Data

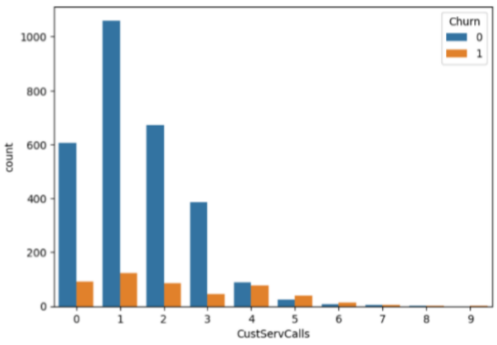
The “Customer Churn” (Kumar, 2020) dataset consists of 3,334 instances and 11 features. The features include ‘Churn,’ ‘AccountWeeks,’ ‘ContractRenewal,’ ‘DataPlan,’ ‘DataUsage,’ ‘CustServCalls,’ ‘DayMins,’ ‘DayCalls,’ ‘MonthlyCharge,’ ‘OverageFee,’ and ‘RoamMins.’

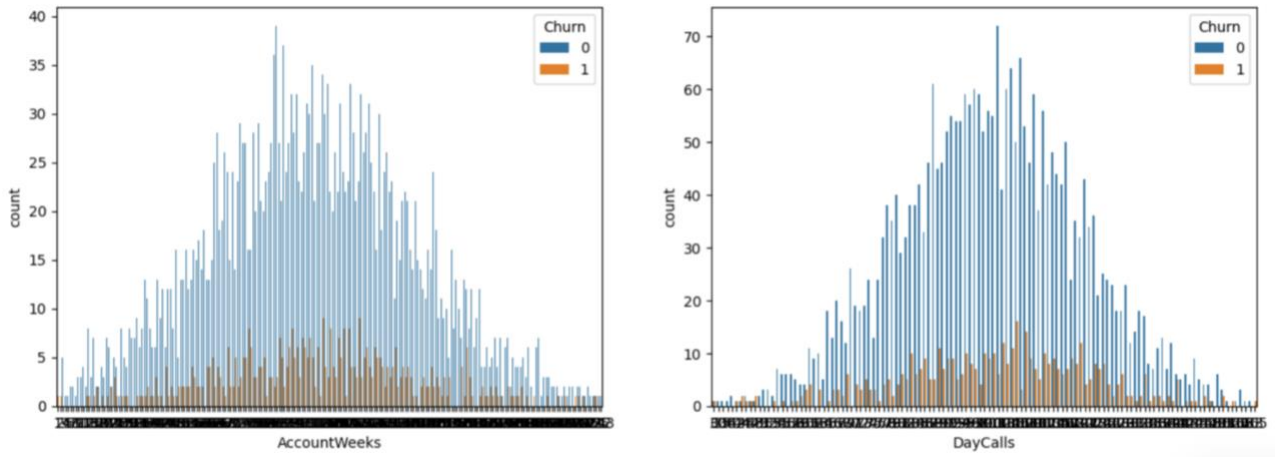
The dataset can be classed as a classification dataset with ‘Churn’ being the target variable and each of ‘AccountWeeks,’ ‘ContractRenewal,’ ‘DataPlan,’ ‘DataUsage,’ ‘CustServCalls,’ ‘DayMins,’ ‘DayCalls,’ ‘MonthlyCharge,’ ‘OverageFee,’ and ‘RoamMins’ being the predictor variables. There are 3 features containing categorical data in the dataset, including the target variable, 3 features containing discrete data, and 5 features containing continuous data.

Bar graphs were created to describe the categorical and discrete features. ‘AccountWeeks’ and ‘DayCalls’ had a relatively normal distribution while ‘CustServCalls’ had a positively skewed distribution.

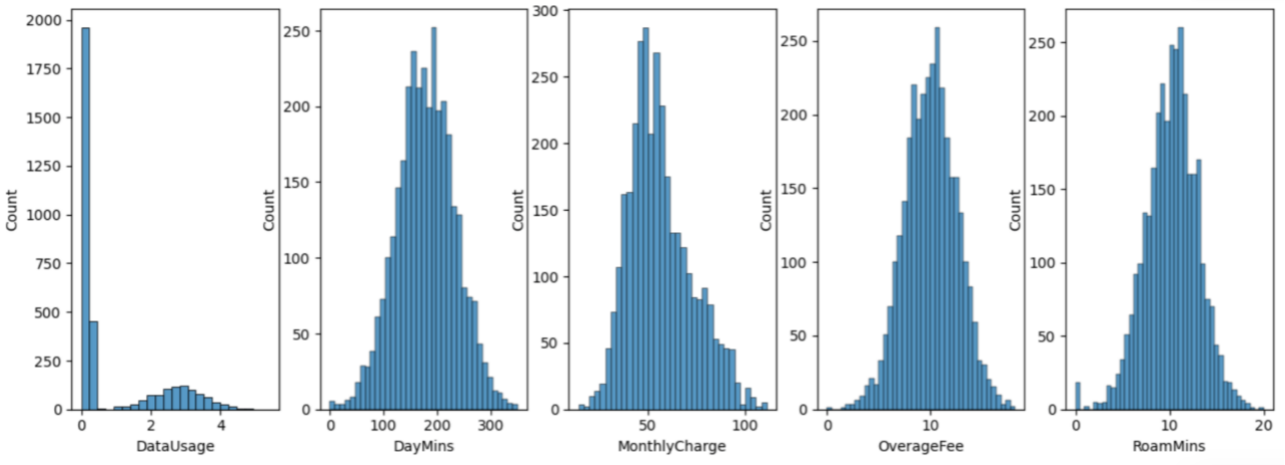






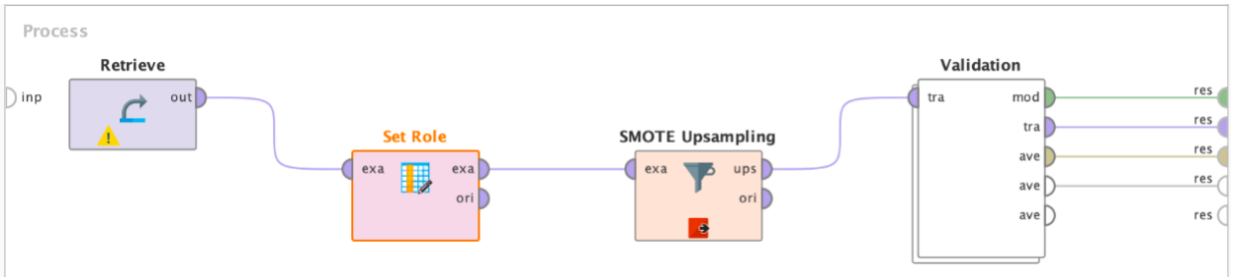


Histograms were created to describe the continuous features. Each of the features had a relatively normal distribution except for ‘DataUsage’ which had a positively skewed distribution.

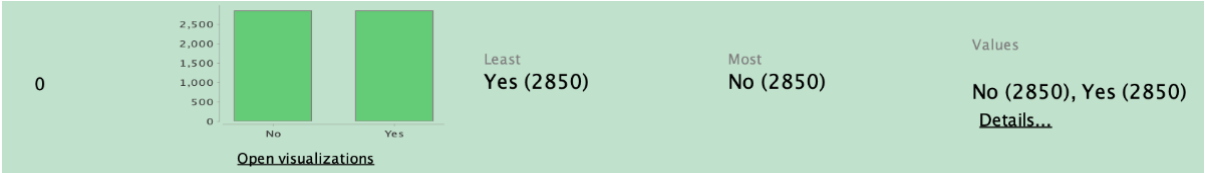


The dataset does not contain any missing or incorrect values but the 0/1 variables in the ‘Churn’ feature were changed to No/Yes variables to allow the Set Role operator to label the variables during the balancing process.

Prior to exploring the potential models that could be used to predict for customer churn the data was balanced in RapidMiner using the process below.



In this process the data is retrieved, the 'Churn' feature is labelled and then the SMOTE Upsampling operator balances the data to prevent overfitting.



Once the data was balanced it was brought to AutoModel to determine which models would give the best scores in order to answer the research question. The results were then recorded in the 'AutoModelling Benchmark' Spreadsheet and can be seen below.

	NB	SD+-	GLM	SD+-	Logit	SD+-	FLM	SD+-	DL	SD+-	DT	SD+-	RF	SD+-	GBT	SD+-	SVM	SD+-
Accuracy	0.81	0.02	0.77	0.02	0.77	0.02	0.77	0.03	0.84	0.02	0.85	0.02	0.85	0.01	0.88	0.01	0.73	0.02
Classification_Error	0.19	0.02	0.23	0.02	0.23	0.02	0.23	0.03	0.16	0.02	0.15	0.02	0.15	0.01	0.12	0.01	0.27	0.02
AUC	0.86	0.02	0.83	0.01	0.83	0.01	0.83	0.01	0.92	0.01	0.88	0.02	0.90	0.01	0.93	0.01	0.82	0.02
Precision	0.83	0.02	0.79	0.03	0.79	0.03	0.79	0.03	0.81	0.02	0.85	0.02	0.86	0.02	0.88	0.02	0.80	0.01
Recall	0.76	0.04	0.73	0.03	0.73	0.03	0.74	0.04	0.89	0.02	0.86	0.03	0.84	0.03	0.87	0.03	0.62	0.04
F_Measure	0.80	0.03	0.76	0.03	0.76	0.03	0.76	0.03	0.85	0.02	0.85	0.02	0.85	0.02	0.88	0.02	0.70	0.03
Sensitivity	0.76	0.04	0.73	0.03	0.73	0.03	0.74	0.04	0.89	0.02	0.86	0.03	0.84	0.03	0.87	0.03	0.62	0.04
Specificity	0.85	0.02	0.81	0.01	0.81	0.01	0.80	0.02	0.79	0.03	0.85	0.01	0.86	0.01	0.88	0.02	0.85	0.03

As seen in the above table 'Decision Trees,' 'Deep Learning,' and 'Gradient Boosted Tree' models came out on top in terms of sensitivity score when predicting customers who churn.

Methodology (Approach) Selection

Potential problem-solving approaches encompassed machine learning algorithms such as decision trees, deep learning, and gradient boosted trees. The selection was bolstered by research conducted on similar datasets, confirming the suitability of our chosen models. Notably, an article on churn prediction by Abdelrahim, Assef & Kaden (2019) for a telecom company utilised decision trees and gradient boosted trees, validating our approach.

These models were deemed appropriate due to their ability to handle non-linear relationships, as evidenced by tree-based models, and to recognise patterns through multiple layers of neurons, characteristic of deep learning. For instance, customer retention decisions may hinge on factors like data usage, wherein individuals with low usage may prioritise other plan features and be happy to stay in the company, while those with high usage may perceive greater value for money, also happy to stay with the company therefore this relationship would be more curved than linear.

Initial results were obtained from Rapid Miner's Automodel, laying a foundational framework for our research. Subsequently, Python on Google Colab was chosen for model application, providing a platform for comprehensive process refinement and increased involvement in model implementation.

Systematic testing of identified approaches involved dataset partitioning into training and testing sets, model fitting, and performance evaluation using metrics like accuracy, precision, recall, and F1 score, with a specific focus on sensitivity. It is worth exploring what each evaluation metric means in relation to the telecommunications dataset:

Accuracy is a measure of how often the model's predictions are correct. It answers the question, "Out of all the customer churn predictions made by the model, how many were accurate?"

Precision is a metric that indicates the proportion of correctly predicted churners out of all the customers predicted to churn by the model. It answers the question, "Out of all the customers predicted to churn by the model, how many actually churned?"

Recall, also known as sensitivity or true positive rate, assesses the model's ability to identify all actual churners correctly. It answers the question, "Out of all the customers who actually churned, how many did the model correctly identify as churners?"

F1 Score is a harmonic mean of precision and recall, providing a balance between the two metrics. It measures the model's overall accuracy in predicting churn while considering the trade-off between precision and recall.

Model Building

Decision trees, deep learning, and gradient boosted trees emerged as the chosen model structures for predicting customer churn.

Each model underwent a standardised process including dataset loading, target variable identification, 70:30 data splitting, dataset balancing, model fitting, and evaluation through accuracy, precision, recall, and F1 score calculations. Emphasis was placed on sensitivity scores, ultimately revealing the deep learning model's superior performance. Emphasising sensitivity in churn prediction is crucial because it directly relates to the model's ability to correctly identify customers who are likely to churn. The scores of these models across the previously mentioned evaluation metrics are shown below:

Decision Trees

	precision	recall	f1-score	support
0	0.93	0.96	0.95	838
1	0.74	0.65	0.70	162
accuracy			0.91	1000
macro avg	0.84	0.81	0.82	1000
weighted avg	0.90	0.91	0.90	1000

Gradient Boosted Trees

	precision	recall	f1-score	support
No	0.95	0.97	0.96	857
Yes	0.78	0.69	0.73	143
accuracy			0.93	1000
macro avg	0.86	0.83	0.84	1000
weighted avg	0.92	0.93	0.93	1000

Deep Learning

	precision	recall	f1-score	support
0	0.97	0.91	0.94	857
1	0.60	0.84	0.70	143
accuracy			0.90	1000
macro avg	0.79	0.87	0.82	1000
weighted avg	0.92	0.90	0.90	1000

Experimentation with different hyperparameters of tree-based models, including tree count and maximum depth, was conducted to enhance sensitivity. Despite adjustments leading to improvements in accuracy and recall, deep learning consistently outperformed tree-based models in sensitivity. The optimal parameter identified for the tree-based models was a maximum depth of 10. Shown below is a line of code that was used to specify the max depth for gradient boosted trees model:

```
# start the Gradient Boosted Trees Classifier
gbt_classifier = GradientBoostingClassifier(max_depth=10)
```

While sensitivity scores were notably high for the deep learning model, they did not surpass those achieved by Rapid Miner's automodel. Although the deep learning model exhibited a higher accuracy rate, the difference in sensitivity scores was minimal, indicating a nuanced trade-off between metrics.

The efficacy of the models is intricately linked to dataset quality and representativeness. Preprocessing efforts included thorough checks for missing values to safeguard data integrity. It is worth noting that deep learning models' black-box nature poses challenges in interpreting decisions, potentially impeding a comprehensive understanding of churn prediction factors (Bagchi, 2022).

Deployment

Deployment is the final stage of the Crisp-DM method. This involves deploying the model that we have built into a production environment. Whether we can deploy the model or not is dependent on the results of our model. If the model does not reach the targeted threshold, then the model would not be deployed and there would be an investigation into why the accuracy of the model was low and use different techniques to improve the accuracy of the model (eg. Hyperparameter tuning). In this case, the model (Deep Learning) has produced satisfactory results on all metrics including accuracy (.90), recall (.87), and sensitivity (.87), therefore the process of deployment can take place.

Deploying the model will involve integrating the model into the business's operational processes. Way of integration include:

Scalability of the Model

Businesses will tend to use larger amounts of data. For that reason, the model needs to be able to handle this increased data volume and user traffic as a business grows. With the use of platforms such as Amazon Web Services (AWS) and Google Cloud, this can help with providing sizable and scalable services for the user.

Ensure User-Friendly Accessibility

Developing an interface that requires training for stakeholders to use the model, this would minimise the need for extensive training. Implement a model with clear visualisations and intuitive controls allowing stakeholders to easily interpret the predictions of the model.

Cost-Efficiency

Using the model for predictions within a company can lead to cost-efficient business decisions. Being able to predict whether a company will stay or not will create benefits for the business such as optimising resource allocation and minimising customer churn related expenses.

Performance Monitoring

Introduce some monitoring mechanisms to track model's effectiveness over time. Get real world feedback and apply to the model to maintain relevance.

Limitations

While the Deep Learning model is the one in which we would plan to deploy, it is important to note some of its limitations or criticism that comes when using said model. Deep learning often relies on large datasets, raising concerns about data privacy and security, especially within a business setting where privacy is of great importance.

Conclusion

The exploration of the “Customer Churn” dataset and subsequent model development was aimed to address the critical business problem of predicting customer churn within a telecom company. With the use of machine learning techniques, particularly using decision trees, gradient boosted trees and deep learning models, our analysis delved into the nuances of customer behaviour and different churn dynamics.

We conducted a comparative analysis between results from our code in Google Colab and RapidMiner’s Automodel feature. This cross-validation approach allowed for us to validate the consistency of our model’s performances.

Through rigorous evaluation and experimentation, our study revealed the deep learning model as the most promising solution to solving customer churn for a telecoms company. This was mainly due to its superior performance in sensitivity, a vital metric for identifying customers who are most likely to churn.

The results obtained from our model, particularly the deep learning approach, exhibited high levels of accuracy, recall and sensitivity, meeting the criteria for a successful deployment. With accuracy metrics reaching 90%, recall 87% and sensitivity reaching 87%, our model demonstrated robust predictive capabilities, providing us confidence in its practical utility.

In conclusion, our comprehensive approach, framing our problem, analytics methodology, model building and deployment strategies, creates a foundation for a telecoms company to look at

insights to enhance customer retention efforts and importantly drive sustainable business growth based on results.

References

Baremetrics (n.d.) “How Churn Prediction Can Improve Your Business” [Online] Available at: https://baremetrics.com/academy/churn-prediction-can-improve-business (Accessed: 19th April 2024)

Dublin Business School (2024) *“Advanced Data & Network Mining: Evaluation”* [Lecture Slides] Available at: https://elearning.dbs.ie/pluginfile.php/2100343/mod_resource/content/0/B9DA110%20-%20Evaluation.pdf (Accessed: 1st April 2024)

Kumar, B. (2020) Customer Churn [Dataset]. Available at: https://www.kaggle.com/datasets/barun2104/telecom-churn/data (Accessed: 1st April 2024)

Larson, J. (2022) *“The Most Revealing Success Metrics for High-Performing Data and Analytics Leaders”* Available at: https://iianalytics.com/community/blog/success-metrics-for-data-and-analytics-leaders (Accessed: 2nd April 2024)

Ranjan, N., Bharambe, Y., Deshmukh, P., Karanjwane, P., and Choudhary, D. (2023) *“Churn Prediction in Telecommunication Industry”* Available at: https://www.researchgate.net/publication/369790574_Churn_Prediction_in_Telecommunication_Industry (Accessed: 19th April 2024).

Abdelrahim, A., Assef, J., & Kadan, A. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*.

Bagchi, S. (2022). What is a black box? A computer scientist explains what it means when the inner workings of AIs are hidden. *The Conversation*.