

Predicting Customer Churn in Telecommunications Using Decision Tree and Random Forest Models

Fouad Bouabdi-Benaatou

Personal Information

Student ID: 2242740
Student Name: Fouad Bouabdi-Benaatou
Course: Artificial Intelligence ACCA7026
Teacher: Mr Malik Uzair Mehmood
Address: University of Wales Trinity Saint David
MSC12 Data and Artificial Intelligence
Birmingham

Contents

1 Introduction 3

2 Related Work 3

3 Methodology 4

3.1 Dataset and data quality 4

3.2 Preprocessing and feature engineering 5

3.3 Models and justification 5

3.4 Hyperparameter tuning 6

3.5 Evaluation and interpretation 6

4 Results and Discussion 6

4.1 Exploratory analysis 6

4.2 Model performance 8

4.3 Overfitting, generalisation and interpretability 9

4.4 Drivers of churn 9

4.5 Profile of customers likely to churn 11

4.6 Bias and fairness 11

4.7 Implications for retention 12

4.8 Suitability and limitations 12

5 Conclusion and Recommendations 12

A Additional Figures 13

B Detailed Model Outputs 15

C Code Excerpts 16

List of Figures 17

List of Tables 17

List of Abbreviations 17

Abstract

Customer churn reduces revenue for telecommunications companies, so operators need to find customers at risk before they leave. This study builds and compares a Decision Tree and a Random Forest classifier on the IBM Telco Customer Churn dataset, which has 7,043 customers and a churn rate of 26.5%. The data were cleaned, encoded and split in a stratified 80/20 way. Both models were tuned with five-fold cross-validation using ROC-AUC, and class weights were used because the classes are imbalanced. On the held-out test set the Random Forest reached a ROC-AUC of 0.845 and the Decision Tree reached 0.836. The difference was small, and a bootstrap interval for it included zero. Permutation importance, SHAP values and partial dependence plots show that contract type, tenure, fibre optic internet and monthly charges are the main drivers of churn. The customers at highest risk are mostly new, on month-to-month contracts, with fibre optic internet and electronic cheque payment. Recommendations and limits are given.

Keywords: customer churn, decision tree, random forest, SHAP, telecommunications, interpretability

1 | Introduction

Customer churn means that a customer stops using the services of a company. In the dataset used here, customers who churned make up 26.5% of all customers but pay 30.5% of the total monthly charges, because their average monthly charge is higher (74.44 against 61.27). The revenue at risk is therefore larger than the customer count suggests.

Machine learning can help by giving each customer a churn score, so a retention team with a limited budget can contact the customers who are most likely to leave. The model can make two kinds of mistake. If it misses a customer who leaves, the company loses that customer without any attempt to keep them. If it flags a customer who would have stayed, effort or discount is wasted. The balance between these two errors is a business decision, and it affects how models should be compared.

This study compares two tree-based classifiers on the IBM Telco Customer Churn dataset. The Decision Tree is simple and can be read as a set of rules. The Random Forest combines many trees and is usually more accurate but harder to explain. The question is whether the extra complexity of the Random Forest gives enough benefit on this data. The objectives are:

- to explore and prepare the dataset;
- to develop and tune both models;
- to compare them using accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices;
- to identify the drivers of churn and describe the customers most likely to leave; and
- to suggest practical retention actions.

The rest of the report reviews related work (Section 2), explains the method (Section 3), presents and discusses the results (Section 4), and gives conclusions and recommendations (Section 5).

2 | Related Work

Churn prediction is normally treated as a binary classification problem. A known difficulty is class imbalance, where one class is much smaller than the other [1]. This applies here, because only 26.5% of customers churn.

Several studies compare classifiers for churn. Vafeiadis et al. compared five techniques, with and without boosting, and the best was a support vector machine with AdaBoost (accuracy almost 97%, F-measure above 84%) [2]. Sabbah compared ten techniques and reported that Random Forest and AdaBoost performed best, with about 96% accuracy [3]. Ahmad et al. used data from the operator SyriaTel. On one of their datasets Random Forest reached an AUC of 83.4% and Decision Tree 79.1%, while XGBoost was best with 89% [4].

More recent studies often add explanation methods. Peng and Peng tuned an XGBoost model with a genetic algorithm and used SHAP to explain it [5]. Chang et al. compared several classifiers, found that Random Forest

performed best, and used SHAP and LIME for explanation [6]. Rasool et al. used SMOTE and compared single and ensemble models, and a stacking ensemble performed best [7].

These studies do not agree on one best algorithm. The best model was a boosted support vector machine, boosted trees, Random Forest or a stacking ensemble, depending on the study. The datasets, preprocessing steps and metrics differ, which is a likely reason, so it is difficult to know how much of each result comes from the algorithm and how much from the setup.

Decision trees learn by splitting the data again and again into smaller groups [8]. A Random Forest combines many trees, each grown with a random element, and its generalisation error depends on the strength of the individual trees and the correlation between them [9]. The cost is interpretability. Lundberg and Lee propose SHAP, which assigns each feature an importance value for a particular prediction [10], and it has been extended to tree-based models [11]. Impurity-based importance from a forest can be biased [12].

Finally, models that support decisions about people can treat groups differently, and Mehrabi et al. survey bias and fairness in machine learning [13]. This report compares the two models on one dataset, with the same split and tuning method.

3 | Methodology

3.1 Dataset and data quality

The file `WA_Telco-Customer-Churn.csv` contains 7,043 customers and 21 columns: demographics, subscribed services, account and billing details, and the target **Churn**. According to the assignment brief, the data are the IBM Telco Customer Churn dataset. An official IBM page for this exact file version could not be confirmed, so the results apply to this file only. There were no duplicate rows and no empty cells except in **TotalCharges**. That column was stored as text and had 11 blank values. All 11 customers had a tenure of zero and none had churned, so they are new customers who had not yet been billed, and the blanks were set to zero. In total, 1,869 customers churned (26.5%).

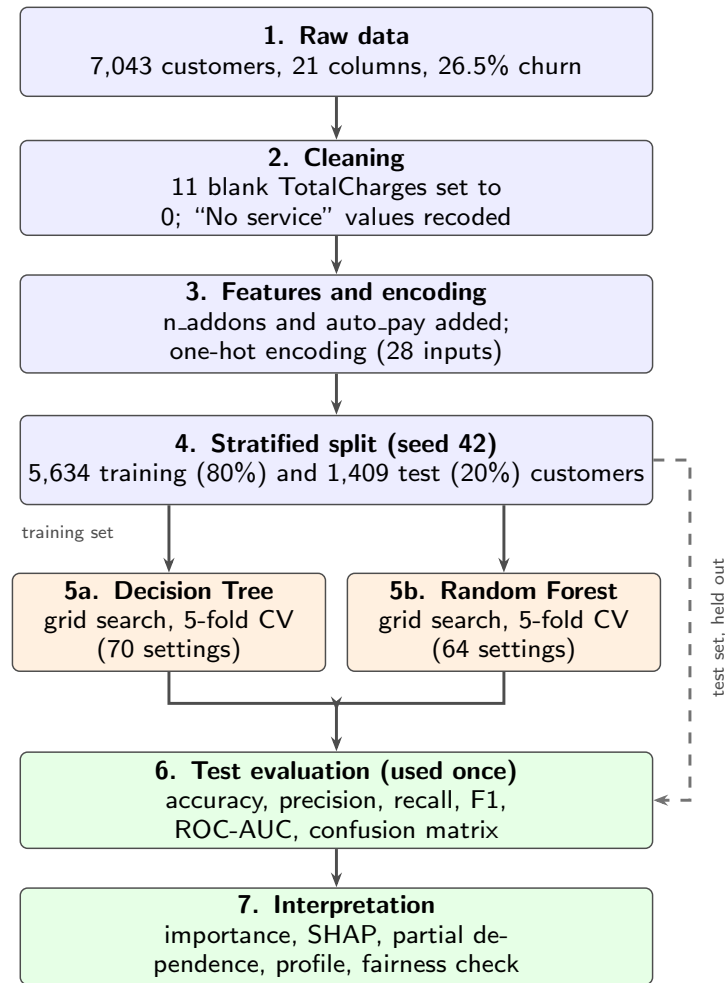


Figure 1: Overview of the modelling workflow.

3.2 Preprocessing and feature engineering

The columns `MultipleLines` and the six internet add-on columns contain the value “No phone service” or “No internet service”. These repeat information already held in `PhoneService` and `InternetService`, so both values were recoded as “No”. Two features were added: `n_addons`, which counts how many of the six add-ons a customer has, and `auto_pay`, which is 1 if the customer pays automatically by bank transfer or credit card. The identifier `customerID` was removed. Categorical variables were one-hot encoded, with one column for binary variables, which gave 28 model inputs. Trees do not need scaling, so numeric variables were left unchanged.

The data were split into 5,634 training and 1,409 test customers, stratified by churn, with a fixed random seed. The churn rate was 26.5% in both parts. The encoder is part of a scikit-learn pipeline [14], so it is fitted on training folds only and no information from the test set is used during preparation.

3.3 Models and justification

A Decision Tree was chosen because its rules can be read and checked by non-technical staff [8]. A Random Forest was chosen because it usually gives better predictions by averaging many trees [9]. Both models used `class_weight="balanced"`, which gives churners more weight to reduce the effect of class imbalance. Resampling methods that create synthetic customers were not used, to keep the training data real and the comparison simple. Untuned models with default settings and no class weights were also trained for reference.

3.4 Hyperparameter tuning

Tuning used grid search with stratified five-fold cross-validation on the training set only, with ROC-AUC as the scoring metric because it does not depend on a single decision threshold. Table 1 shows the search space and the selected values. The Decision Tree search tested 70 combinations and the Random Forest search tested 64.

Table 1: Hyperparameter search space and selected values

Model / parameter	Values searched	Selected
Decision Tree		
criterion	gini, entropy	entropy
max_depth	3, 4, 5, 6, 8, 10, none	5
min_samples_leaf	1, 5, 10, 20, 50	50
Random Forest		
n_estimators	200, 400	200
max_depth	6, 10, 15, none	10
min_samples_leaf	1, 3, 5, 10	10
max_features	sqrt, 0.3	sqrt

3.5 Evaluation and interpretation

The final models were evaluated once on the test set using accuracy, precision, recall, F1-score, ROC-AUC and the confusion matrix, with churn as the positive class and a decision threshold of 0.5. Precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad (1)$$

where TP , FP and FN are the true positives, false positives and false negatives. ROC-AUC is the probability that a randomly chosen churner receives a higher score than a randomly chosen non-churner. To see whether the AUC difference between the models could be due to chance, 1,000 bootstrap samples of the test set were drawn and the AUC difference was calculated for each. Overfitting was checked by comparing training, cross-validation and test results.

To find the drivers of churn, three methods were used on the Random Forest: impurity-based importance, permutation importance on the test set (the drop in AUC when a variable is shuffled) and SHAP values [10, 11] for a random sample of 600 test customers. Partial dependence plots were used for tenure and monthly charges. A customer profile was built from the 10% of test customers with the highest predicted risk, and error rates were compared across gender and senior citizen status. All steps are in the supplied notebook (seed 42, Python 3.13, scikit-learn 1.9.1, SHAP 0.52.0), so the results can be reproduced.

4 | Results and Discussion

4.1 Exploratory analysis

Churn differs strongly between customer groups (Fig. 2). Customers on month-to-month contracts churned at 42.7%, against 11.3% for one-year and 2.8% for two-year contracts. Fibre optic customers churned at 41.9%, compared with 19.0% for DSL and 7.4% for customers without internet. Customers paying by electronic cheque churned at 45.3%. Senior citizens also churned more (41.7% against 23.6%), and customers without online security or technical support churned at about 42%, against about 15% for those who had them. Gender made almost no difference (26.9% and 26.2%).

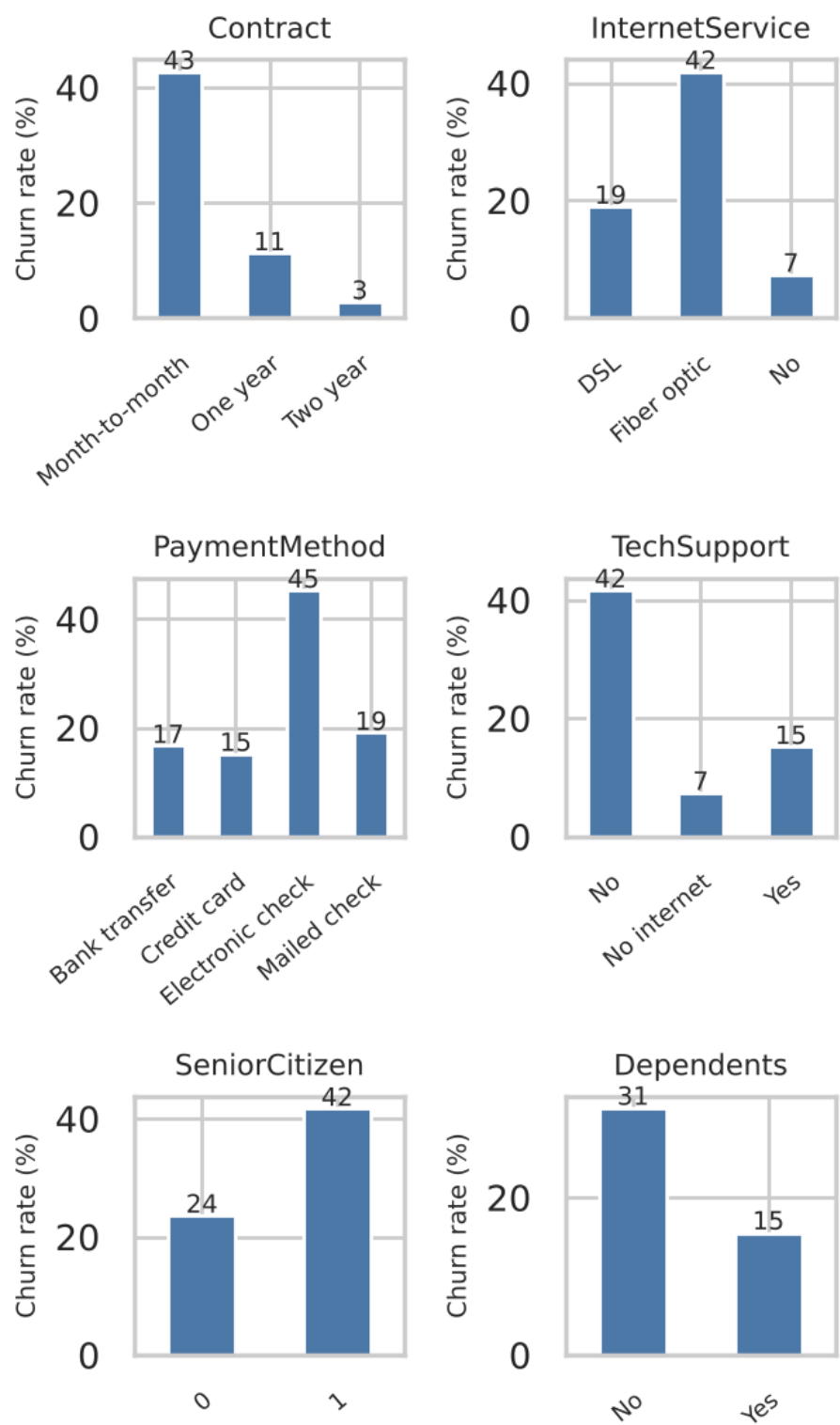


Figure 2: Churn rate (%) by selected customer characteristics. Bank transfer and credit card are automatic payment methods.

Tenure also matters. The median tenure was 10 months for churners and 38 months for customers who stayed,

and 55.5% of churners left within their first 12 months, although only 31.0% of all customers had a tenure of 12 months or less. The correlation analysis agrees: churn has the strongest positive correlation with month-to-month contracts ($r = 0.41$), fibre optic internet (0.31) and electronic cheque payment (0.30), and the strongest negative correlation with tenure (-0.35) and two-year contracts (-0.30). Tenure and total charges are strongly related ($r = 0.83$), which matters when importance scores are interpreted. Further plots are in Appendix A.

4.2 Model performance

Table 2 shows the test results. Tuning had a large effect. The default Decision Tree had a test AUC of 0.667 and recall of 0.527, and it fitted the training data almost perfectly. After tuning, its AUC rose to 0.836 and its recall to 0.783. The default forest was already better (AUC 0.827) but found only 54.0% of churners.

Table 2: Test set performance (1,409 customers, 374 churners)

Model	Acc.	Prec.	Rec.	F1	AUC
Decision Tree (default)	0.732	0.495	0.527	0.510	0.667
Random Forest (default)	0.793	0.629	0.540	0.581	0.827
Decision Tree (tuned)	0.753	0.523	0.783	0.627	0.836
Random Forest (tuned)	0.757	0.528	0.791	0.633	0.845

The tuned models are very close. The Random Forest has a higher AUC (0.845 against 0.836), recall (0.791 against 0.783) and F1-score (0.633 against 0.627). The confusion matrices (Fig. 3) show 296 churners found by the forest and 293 by the tree, out of 374. The difference in AUC is 0.008, and the 95% bootstrap interval is from -0.001 to 0.019 . This interval includes zero, so on this test set it cannot be said that the forest is truly better. Ahmad et al. [4] found a gap of 4.3 AUC points between forest and tree, so the gap here is much smaller.

In practice, the tuned forest flags 561 of the 1,409 test customers (296 true churners and 265 who would have stayed). A retention team using it would contact about 40% of the customer base to reach 79% of the churners. Predicting “no churn” for everyone would score 73.5% on the test set, and the tuned models score 75.3% and 75.7%, only slightly higher, because class weights trade some accuracy for recall. For retention, recall and AUC are therefore more useful than accuracy. Precision is about 0.53 for both tuned models, so about half of the flagged customers would not have left.

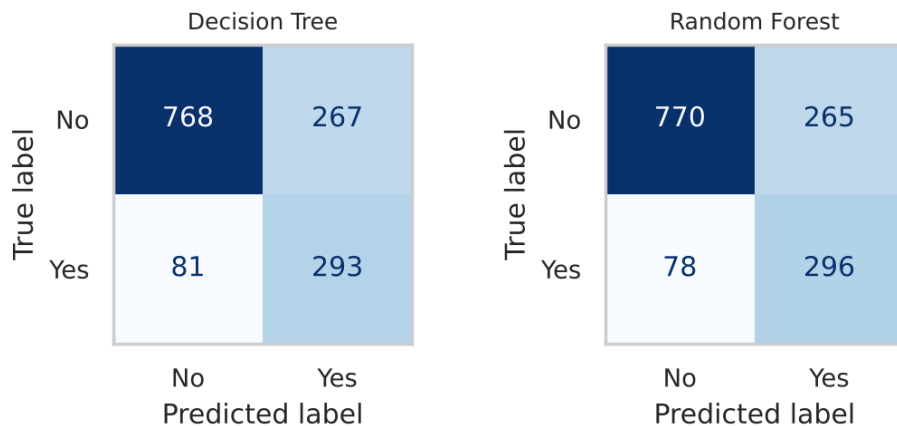


Figure 3: Confusion matrices on the test set for the tuned models.

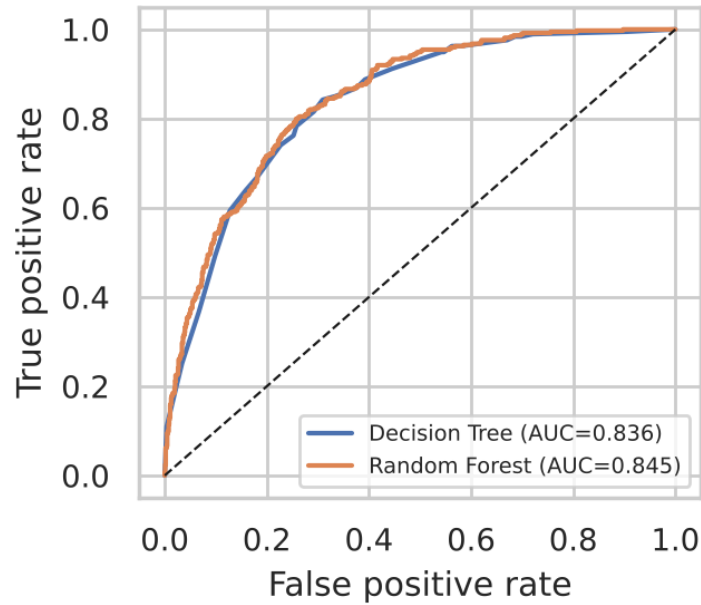


Figure 4: ROC curves on the test set.

4.3 Overfitting, generalisation and interpretability

The tuned models generalise reasonably well. The tuned tree has a training AUC of 0.851 and a test AUC of 0.836, a gap of 0.015. The tuned forest has 0.897 on the training set and 0.845 on the test set, a gap of 0.052. The forest fits the training data more closely, but it also performs better on unseen data, so the extra fit is not harmful here. Before tuning, both default models had a training AUC of 1.000.

The tuned tree has depth 5 and 30 leaves. Its first split is on month-to-month contract, which fits the exploratory results. The top levels are shown as rules in Appendix A. The forest contains 200 trees, which cannot be read directly, so it needs SHAP or similar tools. Given the small difference in AUC, the Decision Tree is a strong choice if staff must understand and explain every decision, and the Random Forest is preferable if the best ranking is the priority.

4.4 Drivers of churn

Figure 5 compares two importance measures. Impurity importance ranks month-to-month contract first (0.191), then tenure (0.149) and total charges (0.099). Permutation importance, calculated on the original variables, gives a different order: contract (AUC drop of 0.085), internet service (0.043) and tenure (0.028), followed by total charges (0.014) and monthly charges (0.005). The lower position of the two charge variables under permutation may reflect the bias in impurity importance [12] and the strong correlation between tenure and total charges, but this was not tested.

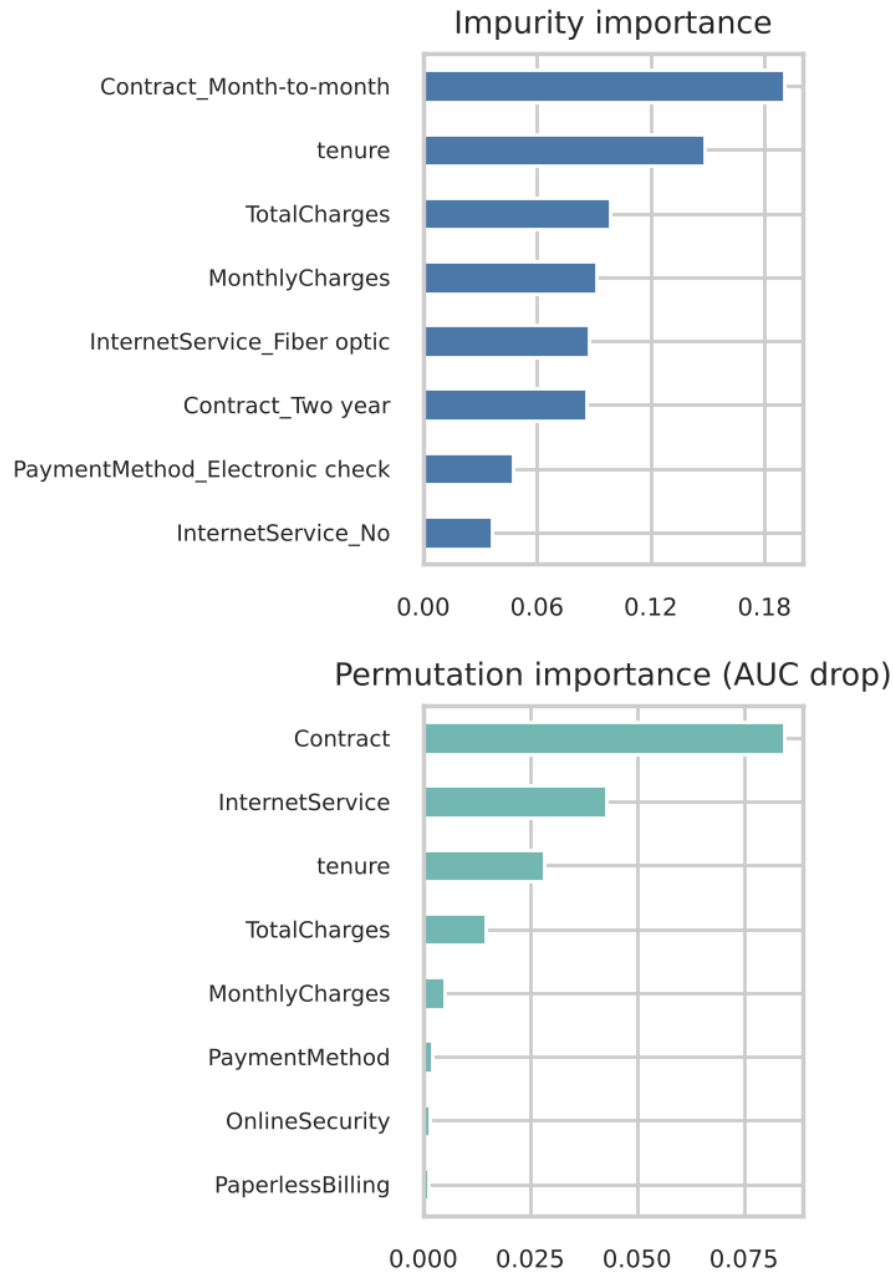


Figure 5: Random Forest feature importance: impurity-based (top) and permutation-based (bottom).

SHAP values (Fig. 6) give the direction of each effect. The largest mean absolute SHAP values were for month-to-month contract (0.091), fibre optic internet (0.066) and tenure (0.062); the top ten are listed in Appendix B. Month-to-month contract, fibre optic internet, electronic cheque payment and high monthly charges push the prediction towards churn, while long tenure, a two-year contract and having no internet service push it away. These results agree with the descriptive statistics, but they show association and not cause.

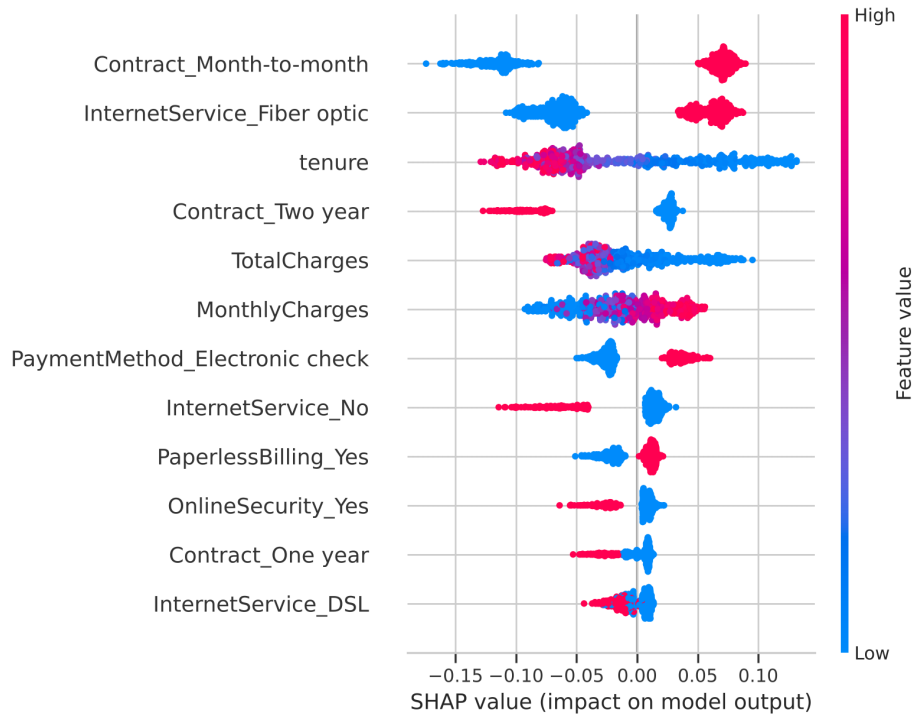


Figure 6: SHAP summary plot for the Random Forest (600 test customers).

The partial dependence plots (Appendix A) show that the predicted churn score falls quickly during the first two years of tenure, from about 0.50 to about 0.38, and more slowly afterwards, to about 0.32 at 72 months. For monthly charges it rises from about 0.35 to about 0.45 above 100. These scores are higher than the real churn rate (26.5%) because class weights shift them, so they are risk scores and not true probabilities.

4.5 Profile of customers likely to churn

The 141 test customers in the top 10% of predicted risk had an actual churn rate of 75.2%, which is 2.8 times the average, and they contained 28.3% of all churners in the test set. All of them had a month-to-month contract. Most had fibre optic internet (95.7%), paid by electronic cheque (76.6%) and used paperless billing (90.1%), and almost none had technical support (6.4%) or online security (1.4%). Their median tenure was 3 months, compared with 33 months for the rest.

4.6 Bias and fairness

The brief mentions fairness in credit scoring. The same issue applies to retention, because the model decides who receives offers. It uses **gender** and **SeniorCitizen** as inputs, so error rates for these groups were compared on the test set (Table 3).

Table 3: Random Forest results by group (test set)

Group	<i>n</i>	Churn	Flagged	Recall	FPR
Female	687	0.281	0.394	0.772	0.247
Male	722	0.251	0.402	0.812	0.264
Not senior	1187	0.233	0.352	0.754	0.231
Senior	222	0.441	0.644	0.898	0.444

For gender, the differences are small: the model flags both groups at similar rates (39.4% and 40.2%), and

recall and false positive rate differ by only a few points. For senior citizens the picture is different. Seniors churn more often (44.1% against 23.3%), so a higher flag rate is expected, but the false positive rate is almost twice as high (44.4% against 23.1%), so many seniors who would have stayed are flagged. The model therefore does not give equal error rates across age groups. The harm is low if the action is a friendly call, but larger if flagged customers receive discounts that others do not. The senior group has only 222 customers, so these figures are uncertain. A model without the gender variable should be tested, since gender had almost no relation to churn.

4.7 Implications for retention

The results suggest three practical points. First, churn risk is concentrated early: more than half of churners leave in the first year, and the highest-risk group has a median tenure of only 3 months, so waiting for signs of leaving may be too late. Second, it is not clear whether longer contracts cause lower churn or whether customers who plan to stay choose them, so a trial with a control group would be needed. Third, the top 10% of predicted risk contains only 28.3% of all churners, so the contact list should go further down the ranking, depending on the budget.

4.8 Suitability and limitations

Both models are suitable for ranking customers by churn risk. The threshold should be set using the cost of an offer and the value of a retained customer, which are not in the dataset. The main limitations are these. The data are a single snapshot, so the models cannot capture how behaviour changes over time. The test set contains only 374 churners, which makes all estimates uncertain. Importance and SHAP results describe the model and not the real causes of churn. Finally, the origin of this version of the dataset could not be verified, so the results may not hold for other operators.

5 | Conclusion and Recommendations

This study compared a tuned Decision Tree and a tuned Random Forest for churn prediction on 7,043 Telco customers. Tuning mattered more than the choice of algorithm: the default Decision Tree reached a test AUC of 0.667, while the tuned tree reached 0.836. The tuned Random Forest had a slightly higher AUC (0.845) and found 296 of 374 churners, against 293 for the tree. The bootstrap interval for the AUC difference included zero, so the evidence for a real advantage is weak. The Decision Tree gives almost the same performance with rules that can be read, while the forest needs SHAP to be explained.

The main drivers of churn were contract type, tenure, internet service type, monthly charges and payment method. The highest-risk customers are new, on month-to-month contracts, with fibre optic internet, electronic cheque payment and no support or security services. These are associations, so the actions below should be tested before wide use.

1. **Focus on new month-to-month customers.** More than half of churners left within 12 months, so early contact and offers of longer contracts should be tested against a control group.
2. **Investigate fibre optic churn.** Churn is 41.9% for fibre optic and 19.0% for DSL. The data do not show why, so price, service quality and competitor offers should be reviewed.
3. **Test support and security bundles.** Customers with technical support or online security churn much less.
4. **Encourage automatic payment.** Electronic cheque customers churn at 45.3%.
5. **Rank customers with the model and set the threshold by cost.**
6. **Monitor fairness.** Track error rates by age group, and test a model without gender.

The analysis uses one dataset and one split, and the fairness check covers only two variables. Future work could test other algorithms and cost-based metrics, and use data that change over time.

A | Additional Figures

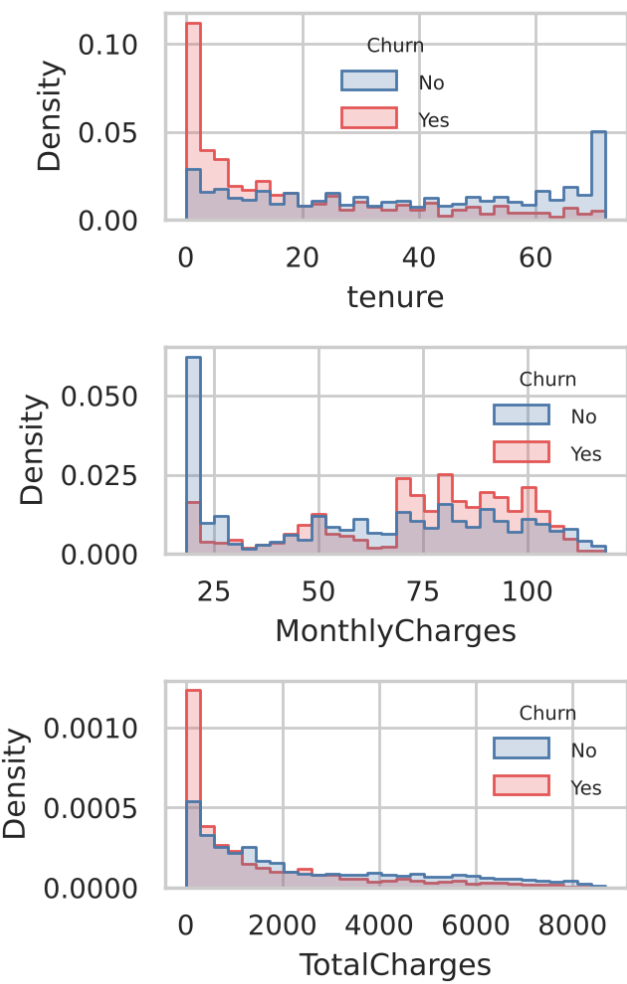


Figure 7: Distribution of tenure, monthly charges and total charges by churn.

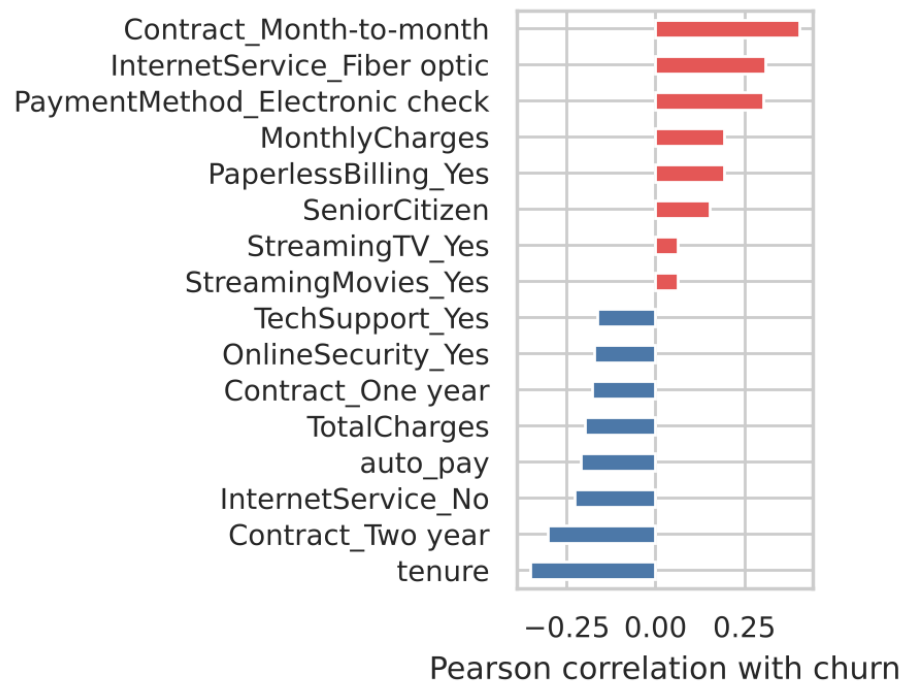


Figure 8: Correlation of encoded features with churn.

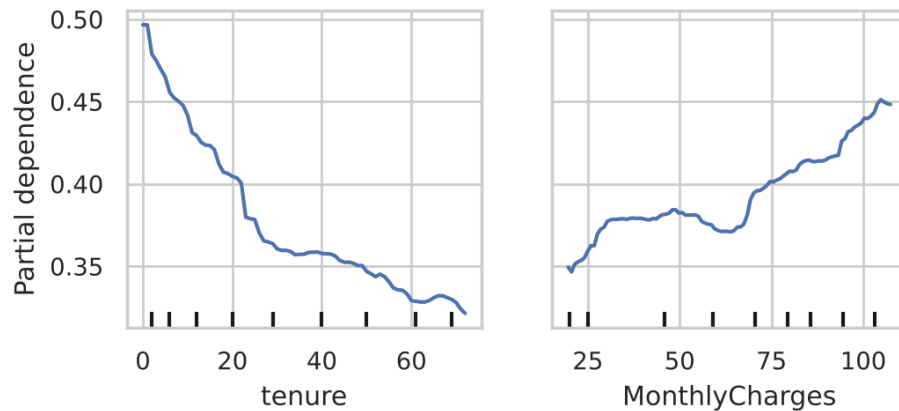


Figure 9: Partial dependence of predicted churn on tenure and monthly charges (Random Forest).

Top two levels of the tuned Decision Tree, as rules. Only the first two levels below the root are shown, and deeper branches are cut.

```

|--- Contract_Month-to-month <= 0.50
| |--- MonthlyCharges <= 93.67
| | |--- Contract_One year <= 0.50
| | | |--- truncated branch of depth 3
| | |--- Contract_One year > 0.50
| | | |--- truncated branch of depth 3
| |--- MonthlyCharges > 93.67
| |--- tenure <= 70.50
| | |--- truncated branch of depth 3
| |--- tenure > 70.50
| | |--- truncated branch of depth 3
|--- Contract_Month-to-month > 0.50

```

```

| |--- InternetService_Fiber optic <= 0.50
| | |--- tenure <= 5.50
| | | |--- truncated branch of depth 3
| | |--- tenure > 5.50
| | | |--- truncated branch of depth 3
| |--- InternetService_Fiber optic > 0.50
| | |--- tenure <= 13.50
| | | |--- truncated branch of depth 3
| | |--- tenure > 13.50
| | | |--- truncated branch of depth 3

```

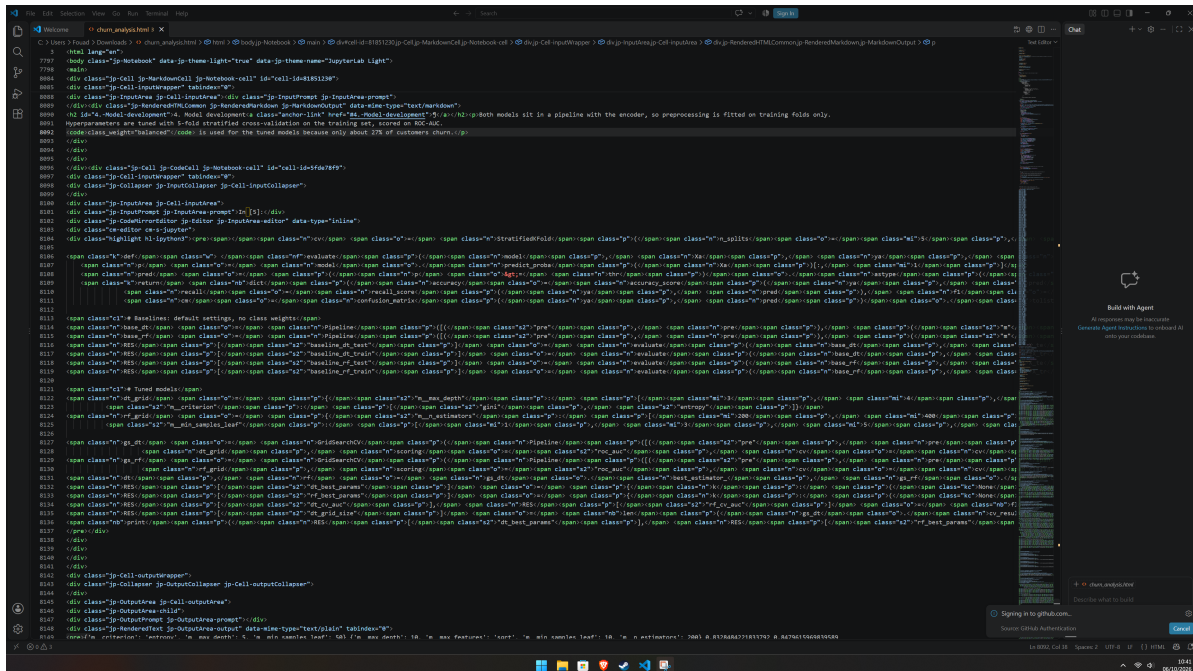


Figure 10: Screenshot of the analysis notebook in Visual Studio Code.

B | Detailed Model Outputs

Table 4: Confusion matrices on the test set (rows: actual, columns: predicted)

Model	TN	FP	FN	TP
Decision Tree (tuned)	768	267	81	293
Random Forest (tuned)	770	265	78	296

Table 5: Random Forest feature rankings (mean absolute SHAP value, 600 test customers)

Feature	Mean SHAP
Contract: month-to-month	0.091
Internet service: fibre optic	0.066
Tenure	0.062
Contract: two year	0.041
Total charges	0.033
Monthly charges	0.031
Payment: electronic cheque	0.029
Internet service: none	0.025
Paperless billing	0.016
Online security	0.014

C | Code Excerpts

The full notebook is supplied as `churn.analysis.ipynb` with an HTML export. The key modelling steps are shown below.

```
pre = ColumnTransformer(
    [("cat", OneHotEncoder(drop="if_binary",
        handle_unknown="ignore", sparse_output=False),
        cat_cols)], remainder="passthrough")
X_tr, X_te, y_tr, y_te = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42)
cv = StratifiedKFold(5, shuffle=True, random_state=42)
gs_rf = GridSearchCV(
    Pipeline([("pre", pre), ("m", RandomForestClassifier(
        class_weight="balanced", random_state=42))]),
    rf_grid, scoring="roc_auc", cv=cv).fit(X_tr, y_tr)
```

List of Figures

Fig. 1. Overview of the modelling workflow	5
Fig. 2. Churn rate (%) by selected customer characteristics	7
Fig. 3. Confusion matrices on the test set for the tuned models	8
Fig. 4. ROC curves on the test set	9
Fig. 5. Random Forest feature importance: impurity-based (top) and permutation-based (bottom)	10
Fig. 6. SHAP summary plot for the Random Forest (600 test customers)	11
Fig. 7. Distribution of tenure, monthly charges and total charges by churn	13
Fig. 8. Correlation of encoded features with churn	14
Fig. 9. Partial dependence of predicted churn on tenure and monthly charges (Random Forest)	14
Fig. 10. Screenshot of the analysis notebook in Visual Studio Code	15

List of Tables

Table 1. Hyperparameter search space and selected values	6
Table 2. Test set performance (1,409 customers, 374 churners)	8
Table 3. Random Forest results by group (test set)	11
Table 4. Confusion matrices on the test set (rows: actual, columns: predicted)	15
Table 5. Random Forest feature rankings (mean absolute SHAP value, 600 test customers)	16

List of Abbreviations

AUC	Area under the curve
CART	Classification and regression trees
CSV	Comma-separated values
CV	Cross-validation
DSL	Digital subscriber line
FN	False negative
FP	False positive
FPR	False positive rate
IBM	International Business Machines
ROC	Receiver operating characteristic
SHAP	Shapley additive explanations
TN	True negative
TP	True positive

References

- [1] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [2] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.
- [3] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018.
- [4] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, no. 1, 2019.
- [5] K. Peng and Y. Peng, "Research on telecom customer churn prediction based on GA-XGBoost and SHAP," *Journal of Computer and Communications*, vol. 10, no. 11, pp. 107–120, 2022.

- [6] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of customer churn behavior in the telecommunication industry using machine learning models," *Algorithms*, vol. 17, no. 6, p. 231, 2024.
- [7] H. A. Rasool, K. K. Aljawaheri, and A. A. M. Karram, "Explainable customer churn prediction in telecom using ensemble learning and SHAP analysis," *Mathematical Modelling of Engineering Problems*, vol. 12, no. 12, pp. 4373–4386, 2025.
- [8] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.
- [11] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020.
- [12] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007.
- [13] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, pp. 1–35, 2021.
- [14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.