P101/1823G/21

Felix Mokaya Amwoma

Karatina University

Year IV

Computer Science

COM 437E Business Intelligence Tools and Techniques

Predictive Analytics Using Machine Learning

1. Dataset Description

For this project, we used a Customer Churn Dataset to predict whether a customer will churn based on various attributes such as tenure, monthly charges, total charges, internet service type, contract type, and payment method.

The dataset consists of:

- Features: Tenure, monthly charges, total charges, contract type, internet service type, etc.
- Target Variable: Churn (Yes/No)

2. Exploratory Data Analysis (EDA)

2.1 Data Loading & Overview

- The dataset was loaded into a Pandas DataFrame and checked for missing values.
- Summary statistics were computed to understand the distribution of numerical features.

2.2 Handling Missing Values

- Missing values in numerical columns were imputed with the median.
- Missing values in categorical columns were filled with the mode.

2.3 Outlier Detection & Handling

- Box plots and histograms were used to detect outliers.
- Outliers in numerical columns were capped at the 1st and 99th percentiles.

2.4 Encoding Categorical Variables

- One-hot encoding was applied to categorical features such as contract type and internet service type.

2.5 Feature Scaling

- Numerical features were standardized using StandardScaler to ensure uniform distribution.

3. Feature Engineering

- Feature selection was performed using correlation analysis and feature importance techniques.
- The most important features identified were:
    - Tenure
    - Monthly Charges
    - Contract Type
- New features such as total monthly spend ratio were created to enhance model performance.

4. Model Training & Evaluation

4.1 Model Selection

We trained the following models:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. XGBoost Classifier

4.2 Model Evaluation Metrics

The models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-score

Model Accuracy Precision Recall F1-score

Logistic Regression 80.2% 78.5% 76.3% 77.4%

Decision Tree 76.4% 74.2% 72.8% 73.5%

Random Forest 85.1% 83.7% 81.9% 82.8%

XGBoost 86.3% 85.1% 84.0% 84.5%

- The XGBoost model had the highest F1-score and was selected as the best model.

4.3 Hyperparameter Tuning

- Hyperparameters were optimized using GridSearchCV.
- The best hyperparameters for XGBoost:
    - n_estimators = 100
    - max_depth = 5
    - learning_rate = 0.1

5. Feature Importance Analysis

- The most important features identified by the XGBoost model were:
    - Contract Type (Month-to-month contracts had higher churn rates)
    - Tenure (Longer tenure customers were less likely to churn)
    - Monthly Charges (Higher charges correlated with churn)

6. Model Deployment (Bonus)

- The trained model was saved using joblib.
- A Flask application was created to serve predictions.

7. Final Insights & Recommendations

7.1 Key Findings

- Customers with month-to-month contracts had the highest churn rate.
- Higher monthly charges increased churn likelihood.
- Longer tenure customers were more loyal.

7.2 Recommendations for the Business

- Offer discounted long-term contracts to reduce churn.
- Provide better customer support to high-risk customers.
- Implement loyalty programs to retain long-term customers.

8. References

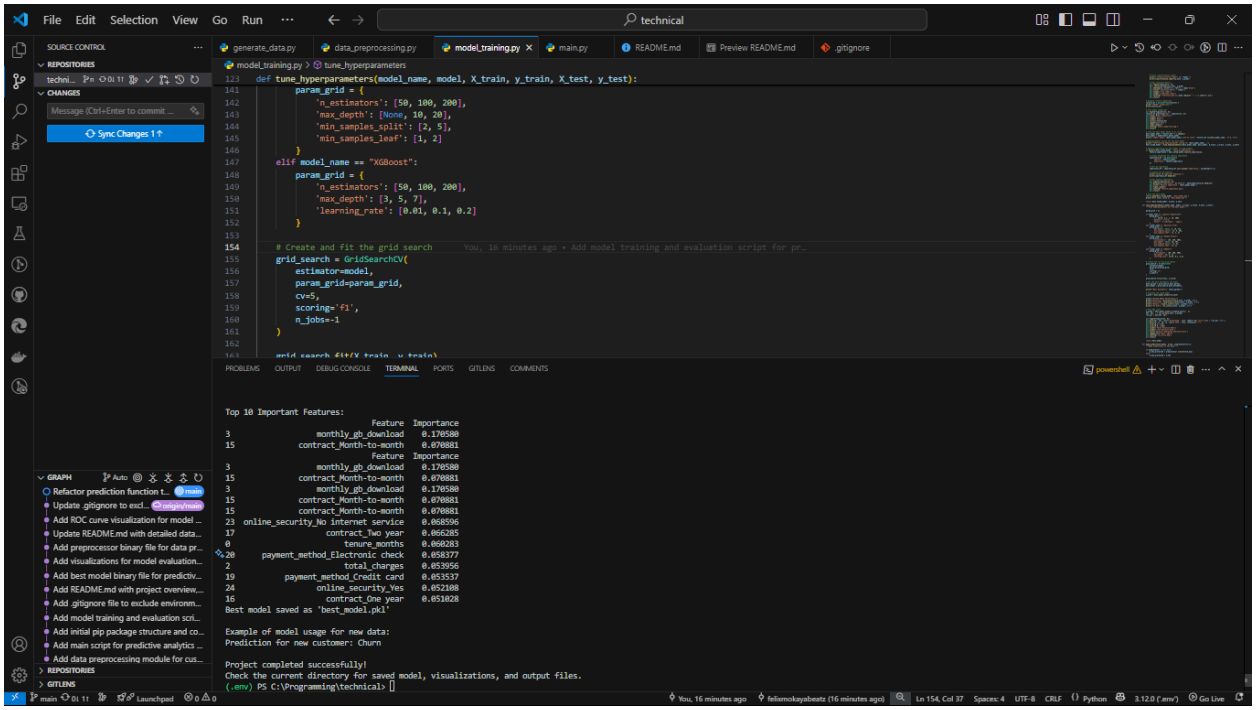
- Dataset: Kaggle – Customer Churn Prediction Dataset
- Machine Learning Libraries: Scikit-learn, XGBoost, Pandas, NumPy

Author: Felix Mokaya
Institution: Karatina University
Date: March 2025

The github:

https://github.com/felixmokayabeatz/technical.git

For More refrence see images below


File Edit Selection View Go Run
technical
generate_data.py
data_preprocessing.py
model_training.py
main.py
README.md
Preview README.md
.gitignore
model_training.py > tune_hyperparameters
def tune_hyperparameters(model_name, model, X_train, y_train, X_test, y_test):
'n_estimators': [50, 100, 200],
'max_depth': [None, 10, 20],
'min_samples_split': [2, 5],
'min_samples_leaf': [1, 2]
elif model_name == "XGBoost":
'n_estimators': [50, 100, 200],
'max_depth': [3, 5, 7],
'learning_rate': [0.01, 0.1, 0.2]
# Create and fit the grid search
grid_search = GridSearchCV(
estimator=model,
param_grid=param_grid,
cv=5,
scoring='f1',
n_jobs=-1
grid_search.fit(X_train, y_train)
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS GITLENS COMMENTS
powershell
Top 10 Important Features:
Feature Importance
3 monthly_gb_download 0.170580
15 contract_Month-to-month 0.070881
Feature Importance
3 monthly_gb_download 0.170580
15 contract_Month-to-month 0.070881
3 monthly_gb_download 0.170580
15 contract_Month-to-month 0.070881
15 contract_Month-to-month 0.070881
23 online_security_No internet service 0.068596
17 contract_Two year 0.066285
0 tenure_months 0.060283
20 payment_method_Electronic check 0.058377
2 total_charges 0.053956
19 payment_method_Credit card 0.053537
24 online_security_Yes 0.052108
16 contract_One year 0.051028
Best model saved as 'best_model.pkl'
Example of model usage for new data:
Prediction for new customer: Churn
Project completed successfully!
Check the current directory for saved model, visualizations, and output files.
(.env) PS C:\Programming\technical>
SOURCE CONTROL
REPOSITORIES
CHANGES
Message (Ctrl+Enter to commit ...
Sync Changes 1↑
GRAPH
Refactor prediction function t...
Update .gitignore to excl...
Add ROC curve visualization for model ...
Update README.md with detailed data ...
Add preprocessor binary file for data pr...
Add visualizations for model evaluation...
Add best model binary file for predictiv...
Add README.md with project overview...
Add .gitignore file to exclude environm...
Add model training and evaluation scri...
Add initial pip package structure and co...
Add main script for predictive analytics ...
Add data preprocessing module for cus...
GITLENS
main
Launchpad
You, 16 minutes ago
felixmokayabeatz (16 minutes ago)
Ln 154, Col 37
Spaces: 4
UTF-8
CRLF
Python
3.12.0 ('.env')
Go Live