CUSTOMER CHURN PREDICTION USING MACHINE LEARNING IN DATAIKU

ABOUT THE PROJECT

Customer churn prediction is a critical aspect of customer relationship management, aiming to identify customers who are likely to stop using a service. Retaining existing customers is widely recognized to be considerably more cost-effective than acquiring new ones, with studies suggesting it can be five to twenty-five times cheaper. As a result, predicting churn has become a strategic priority for organizations aiming to reduce revenue loss and improve customer satisfaction. The project utilizes Dataiku, a data science platform for the seamless integration of data preparation, modeling, and deployment. Dataiku was used to import and clean the dataset, perform exploratory data analysis, and engineer relevant features necessary for effective modeling. With its visual workflows and code integration capabilities, machine learning algorithms such as Random Forest and K-Means clustering were applied within the platform to classify or group customers based on their likelihood to churn. Finally, the results were exported and analyzed to interpret key customer behavior patterns contributing to churn, thereby enabling the development of data-driven strategies for customer retention.

PROBLEM STATEMENT

A leading telecommunications company is experiencing a significant challenge with customer churn. In the past few months, the churn rate has increased leading to a significant loss in the annual recurring revenue. The company currently lacks a proactive mechanism to identify customers at high risk of churning before they actually leave. Existing reactive measures, such as win-back campaigns, are often initiated too late and have shown limited success. This inability to anticipate churn hinders the company's efforts to implement targeted retention strategies, resulting in a continuous drain on its customer base and profitability.Therefore, there is an urgent need to develop a robust and accurate customer churn prediction model.

OBJECTIVES

- **To perform comprehensive exploratory data analysis (EDA)** in order to understand the underlying customer behavior, identify key trends, usage patterns, and uncover significant factors contributing to customer churn.To clean, preprocess, and engineer features from the raw dataset for optimal machine learning model performance.

- **To clean, preprocess, and engineer features from the raw telecom dataset** by handling missing values, correcting formatting issues, transforming categorical variables, and creating meaningful derived features that enhance model interpretability and performance.

- **To build and evaluate a machine learning model using the Random Forest algorithm** for churn prediction. The objective is to accurately identify customers who are at high risk of leaving, enabling proactive engagement and retention actions.

- **To segment the customer base using K-Means Clustering**, allowing the identification of distinct customer groups based on their usage metrics and behavior. This enables the development of targeted marketing strategies and personalized retention plans tailored to each segment.

- **To visualize analytical results and machine learning insights** through interactive dashboards, charts, and flow views within the Dataiku platform, promoting data-driven decision-making and enhancing stakeholder understanding of churn dynamics.

- **To minimize revenue loss and improve customer lifetime value (CLV)** by combining predictive analytics and segmentation insights to focus resources on the most impactful customer groups.

TOOLS & TECHNOLOGIES

- Platform: Dataiku
- **Dataset Source**: Telecom Churn Dataset, downloaded from Kaggle *(https://www.kaggle.com/datasets/mnassrib/telecom-churn-datasets)*
- Language: Python (Dataiku's Python Recipe for K-Means Clustering Model)
- ML Algorithms: K-Means, Random Forest
- Libraries: pandas, numpy, scikit-learn

DATASET DESCRIPTION

- The dataset was sourced from Kaggle and consisted of two CSV files containing telecom customer data.
- Both CSV files had the same structure and fields; they were unified into a single dataset using Join Recipe for analysis.
- The dataset included customer-level information such as demographics, account details, service usage, and a target variable called Churn, indicating whether a customer left the service.
- After merging, data cleaning steps were performed, including removal of outliers, filling missing values, and renaming inconsistent column headers.
- Dataiku's Prepare recipes were used to perform these preprocessing tasks efficiently in a visual and low-code format.
- Exploratory Data Analysis (EDA) was conducted using charts to uncover trends and correlations between features and customer churn.
- Patterns such as service type, contract duration, and international services were found to be related to churn behavior.

METHODOLOGY

- **Data Collection:** The dataset was sourced from Kaggle and comprised two CSV files with similar fields. These were joined into a single unified dataset for analysis.
- **Data Inspection:** The target column for prediction was identified as Churn, a Boolean field. The dataset contained features such as total call minutes, call charges, number of calls, and customer service interactions.
- **Data Cleaning & Preparation:** Used Dataiku's Prepare Recipes to rename columns, eliminate outliers, and fill missing values. This also included data type corrections and formatting.
- **Exploratory Data Analysis (EDA):** Visual tools like bar and pie charts were used to explore trends and identify dependencies between churn and various customer behavior metrics.
- **K-Means Clustering:** A Python recipe was used to manually write and apply a K-Means clustering model to group similar customers based on selected features.
- **Feature Engineering:** The resulting cluster labels were appended to the dataset to enhance model input and improve predictive performance.
- **Random Forest Classification:** Built and trained a Random Forest model using the clustered dataset to classify customer churn.
- **Model Evaluation:** Evaluated performance using accuracy, confusion matrix, and other classification metrics. The best-performing model was selected for prediction.
- **Prediction & Output:** The final model was used to predict churn on test data, and predictions were exported for further analysis, thereby estimating revenue loss.

DATAIKU RECIPES USED

- **Prepare Recipe**

1. **Data Cleaning and Initial Processing -** Prepare recipe was used to clean the datasets at the initial stage, by removing the outliers, filling the empty rows, renaming columns and creating new columns.
2. **Feature Engineering -** For estimating the potential revenue loss, a new column was created using formula, hence Prepare recipe was used for that.

- **Split Recipe**
    1. Using Split recipe, the created KMeans model object dataset was randomly dispatched into two datasets for training and testing.

- **Python Recipe**
    1. For drafting a KMeans Clustering model, a python recipe was used to write the code for the same.

- **Score Recipe**
    1. Score recipe aids in the practical implementation of trained machine learning models. It used to introduce the trained model to new unseen data for making predictions.

PARAMETERS ANALYSED

- Dashboard 1 - <u>Descriptive Statistics on the Customers</u>

    1. This dashboard presents a visual and analytical overview of customer churn patterns in a telecommunications dataset. It leverages a variety of charts and visual tools to uncover factors influencing customer churn and to support data-driven decision-making.
    2. The **churn rate distribution** is represented using a pie chart, showing that approximately **14.6% of customers have churned**, while **85.4% have remained loyal**. A deeper categorical breakdown by **international plan**

usage reveals that customers **with an international plan are significantly more likely to churn** compared to those without it — suggesting dissatisfaction or cost sensitivity related to the plan.

3. A **histogram of total calls** displays usage ranges and highlights that churners mostly fall within the **moderate usage range (200–400 calls)**, indicating that very low or very high usage isn't necessarily the key churn driver. Further, a **box plot analyzing customer service calls** reveals that churned customers tend to have made **more customer service contacts**, potentially signaling unresolved issues or frustration leading to churn.
4. To enhance pattern detection, a **3D scatter plot** is included, mapping total calls, day charges, and churn status. This visualization reveals that **churned customers tend to be concentrated in higher day charge zones**, pointing toward pricing as a potential churn factor.
5. To enhance pattern detection, a **3D scatter plot** is included, mapping total calls, day charges, and churn status. This visualization reveals that **churned customers tend to be concentrated in higher day charge zones**, pointing toward pricing as a potential churn factor. Overall, the dashboard offers actionable insights: churn is influenced by international plan enrollment, frequency of customer service interactions, and day-time billing. These factors can be targeted for customer retention strategies, such as improved service support and pricing revisions.

- Dashboard 2 - <u>Segments Analysis</u>

1. This dashboard visualizes the segmentation of customers using the **K-means clustering algorithm**, providing a clear overview of customer behavior patterns and churn risk based on usage metrics and service interactions.The **segment distribution** is shown using a donut chart, where the total customer base is divided into **three clusters**:
    1) Cluster 0: 4404 customers
    2) Cluster 1: 1830 customers
    3) Cluster 2: 4174 customers

2. The **Churners/Segments bar chart** reveals that churners are not evenly distributed. Cluster 1 has a **higher proportion of churners** compared to clusters 0 and 2, suggesting that this segment includes more dissatisfied or high-risk users.

3. A **feature importance bar chart** ranks the variables influencing cluster assignment. Key contributing factors include total_minutes, day_minutes, churn. These metrics indicate that overall usage and day-time calling activity play major roles in segment differentiation.

4. The **scatter plot** compares day calls with day charges, colored by cluster, helping identify visual separation between user types. Clustered points reveal clear groupings based on usage intensity and billing patterns, supporting the model's segmentation quality.

5. Finally, **individual cluster profiles** are displayed using histograms for each cluster based on the number of customer service calls. These plots illustrate:
   Cluster 0: Low to moderate customer service interaction (median = 1)
   Cluster 1: Slightly higher and more spread-out interactions (median = 102 day calls)
   Cluster 2: Similar to Cluster 1 but with a wider distribution in call behavior (median = 100 day calls)

6. This analysis allows stakeholders to tailor strategies for each cluster — improving retention efforts for high-risk groups, optimizing plans for heavy users, and maintaining satisfaction for stable segments.

- Dashboard 3 - Forecasting Customers that are likely to churn

    1. This dashboard project presents a detailed analysis of customer churn behavior through data visualization, clustering, and machine learning models. The dashboard integrates key insights into customer characteristics, churn distribution, and segment-specific behaviors, providing actionable insights to reduce churn and improve customer retention.

    2. The dashboard begins with a churn overview showing that a significant majority of customers have not churned. A pie chart indicates that out of 13,009 customers, 1,901 (about 15%) have churned. A churn analysis table segmented by international plan usage reveals that customers with an international plan show a higher churn rate than those without. Additionally, histograms display correlations between total calls and churn behavior, emphasizing that churned customers often have higher total call volumes. A box plot and 3D scatter plot further elaborate this relationship with service calls, call volume, and charges, illustrating patterns between customer service usage and churn likelihood.

    3. The segmentation section employs the K-means clustering model to group customers into three distinct clusters based on their usage patterns. Segment distribution shows three clusters of sizes 4404, 4174, and 1830 respectively. A bar chart mapping churn rates per cluster reveals that churn behavior is uneven across segments, with clusters 0 and 2 showing higher churn counts than cluster 1.

4. Variable importance charts highlight that total_minutes, day_minutes, and churn status are among the most impactful features in defining clusters. A scatter plot between day calls and charges visualizes the cluster dispersion. Additionally, cluster profiles show feature-wise distributions, such as customer service calls and minutes, with cluster-specific standard deviations and medians. These insights enable targeted marketing and service improvement strategies for each customer group.

5. The final section introduces a machine learning model using the Random Forest algorithm to predict churn. The model's decision tree visualizes key splits, including thresholds on international calls, day charges, and voicemail usage. The model assigns the highest churn probability to customers with higher international usage and lower voicemail activity.

6. A variable importance chart confirms that day minutes, total minutes, and day charges are the top predictors. Performance metrics including Accuracy (99%), Precision (98%), Recall (95%), and F1-score (96%) indicate excellent model performance. The confusion matrix confirms low misclassification rates, while the cost matrix evaluates the financial benefit of the model's predictions, reporting an average gain per record of 0.14.

7. A revenue loss histogram shows the relationship between clusters and financial impact, emphasizing the importance of identifying at-risk segments (especially cluster 0) to prevent further revenue drain.

CONCLUSION

The project successfully predicted customer churn using K-Means for segmentation and Random Forest for classification in Dataiku. Key churn indicators included customer service

calls and total minutes. Combining clustering with supervised learning improved accuracy and an estimated revenue loss was calculated for predicted churn customers.