**University of Petroleum and Energy Studies**

**Internship - High Level Design**

**on**

**Predicting Customer Churn in Telecommunication Company**

Team members:

Jitender Dhariwal - R2142201723

Jatin Wadhwa - R214220562

Kartikay Tyagi -R214220603

Jatin Joshi - R214220559

Karan Taneja - R214220587

Guided by: Mr. Sumit Shukla

Industry Mentor: Mr. Sumit Shukla

Table of Contents

1. Introduction............................................................................................................................... 6
   1. Scope of the document............................................................................................................................. 6
   2. Intended Audience................................................................................................................................... 6
   3. System overview...................................................................................................................................... 6
2. System Design............................................................................................................................ 6
   1. Application Design................................................................................................................................. 6
   2. Process Flow........................................................................................................................................... 6
   3. Information Flow.................................................................................................................................... 7
   4. Components Design................................................................................................................................ 8
   5. Key Design Considerations…................................................................................................................. 8
   6. API Catalogue.......................................................................................................................................... 8
3. Data Design............................................................................................................................... 8
   1. Data Model............................................................................................................................................. 8
   2. Data Access Mechanism......................................................................................................................... 9
   3. Data Retention Policies.......................................................................................................................... 9
   4. Data Migration........................................................................................................................................ 9
4. Interfaces………………………………….……………………………………………………9
5. State and Session Management.................................................................................................10
6. Caching……………………………………………………………………………………….10
7. Non-Functional Requirements…….........................................................................................10
   1. Security Aspects.................................................................................................................................10
   2. Performance Aspects.........................................................................................................................10
8. References.................................................................................................................................11

**1. Introduction**

For telecommunications firms, client churn—the process in which consumers transfer from one provider to another—raises serious concerns. Retaining clients has risen at the top of many companies' priorities due to fierce competition and changing market conditions. Telecommunications businesses are increasingly using machine learning techniques to forecast client attrition in an effort to reduce the negative effects of churn.

Machine learning models provide an effective way to examine vast amounts of consumer data and spot trends that may be used to identify which customers are most likely to leave. These models can assist telecoms firms take preventative steps to retain at-risk consumers by making use of past customer information.

In the telecoms sector, the aim of customer churn prediction using machine learning is to create models that can precisely identify consumers who are likely to churn in the near future. Companies can enhance customer happiness and loyalty by recognizing these clients in advance and taking the required actions, such as making personalized retention offers, performing outstanding customer service, or launching focused marketing efforts.

Telecommunications firms are able to more efficiently manage their resources because to the predictive capabilities of machine learning algorithms. These models help businesses to concentrate their efforts on consumers who are most likely to churn, allowing them to optimize their retention tactics and lower the expenses involved with gaining new customers.

Furthermore, customer churn prediction models can provide insights into the factors that drive customer attrition. By analyzing the most important features contributing to churn, such as poor network quality, pricing dissatisfaction, or competitor offers, companies can gain a deeper understanding of customer preferences and pain points. This information can guide strategic decision-making, product development, and customer experience improvements.

In this era of data-driven decision-making, customer churn prediction using machine learning offers telecommunications companies a competitive advantage. By harnessing the power of advanced analytics, these companies can enhance customer retention efforts, increase customer satisfaction, and ultimately improve their bottom line.

In this guide, we will explore the steps involved in building a machine learning model for customer churn prediction in a telecommunications company. We will discuss data collection, pre-processing, feature engineering, model selection, training, evaluation, and deployment. By following these steps, telecommunications companies can develop robust churn prediction models and make data-driven decision.

**Scope of the document**

The different stakeholders within a telecom firm who can gain from the insights and forecasts produced by the churn prediction system make up the intended audience of customer churn prediction in the telecom industry. These parties play a role in business strategy, customer relationship management, and decision-making. Key target audiences include the following:

Marketing and sales teams can use churn forecasts to create targeted marketing campaigns and promotional offers to keep hold of high-value clients. Cross-selling and upselling opportunities can be found with the use of churn projections.

Customer Retention Teams: Customer retention teams are essential to the implementation of churn-reduction measures. Churn projections give them useful information about high-risk clients, enabling them to concentrate on tailored retention initiatives.

Customer support teams can use churn projections to proactively solve customer issues and deliver better service, ultimately increasing customer satisfaction and loyalty.

Product Development Teams: Churn forecasts can help product development teams to better understand customer preferences and pain areas so they can design services that better satisfy those requirements and expectations.

corporate analysts: To better understand customer behavior, spot trends, and offer data-driven advice to improve corporate performance, business analysts examine churn prediction data.

Executives and Management: In order to make strategic decisions, distribute resources effectively, and promote long-term business success, a telecom company's senior management depends on churn estimates.

**Intended audience**

The intended audience of customer churn prediction in the telecom industry includes various stakeholders within a telecom company who can benefit from the insights and predictions generated by the churn prediction system. These stakeholders are involved in decision-making, customer relationship management, and business strategy. Some of the key audience groups are:

Marketing and Sales Teams: Marketing and sales teams can leverage churn predictions to develop targeted marketing campaigns and promotional offers to retain high-value customers. Churn predictions can also help in identifying opportunities for cross-selling and upselling services.

Customer Retention Teams: Customer retention teams play a critical role in implementing strategies to reduce churn. Churn predictions provide them with valuable insights on high-risk customers, allowing them to focus on personalized retention efforts.

Customer Support and Service Teams: Customer support teams can utilize churn predictions to proactively address customer issues and provide better service, ultimately enhancing customer satisfaction and loyalty.

Product Development Teams: Churn predictions can offer insights into customer preferences and pain points, guiding product development teams to design services that better meet customer needs and expectations.

Business Analysts: Business analysts analyze churn prediction data to understand customer behavior, identify trends, and make data-driven recommendations to improve business performance.

Executives and Management: The top management of a telecom company relies on churn predictions to make strategic decisions, allocate resources efficiently, and drive long-term business growth.

**System overview**

A telecommunications business can use the churn prediction model to identify customers who are likely to abandon their services in the near future. The organization can use targeted retention measures to lower customer churn and boost customer loyalty by proactively identifying potential churners. Obtaining historical client data from a variety of sources, including user profiles, call logs, payment details, usage statistics, and customer service interactions, is the first stage. In order to handle missing values, normalize the data, and engineer features, this data is then preprocessed. To build the model, pertinent features are chosen from the preprocessed data. These characteristics could include things like client demographics, usage habits, tenure, contract type, call length, complaints from customers, and more. In the selection process, qualities that are instructive are prioritized over those that are unnecessary or redundant. The chosen features are trained using past customer data by a machine learning model. For churn prediction, methods including logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting, and neural networks are frequently utilized. To properly assess the performance of the model, the dataset is divided into training and validation sets. Metrics including accuracy, precision, recall, F1-score, and ROC-AUC are used to assess the model's performance. To guarantee the model's generalizability and avoid overfitting, cross-validation techniques may be used. Following model training, a cutoff point is selected to categorize consumers as churners or non-churners based on the anticipated likelihood scores. According on the company's priorities, the threshold number can be changed to create a balance between false positives and false negatives. To process real-time customer data, the trained churn prediction model is incorporated into the telecom company's current systems. The model forecasts the likelihood of customer churn for specific customers as fresh data comes in. Customer segmentation can also be done using the churn prediction model. Based on customer turnover probability, customers are divided into various segments, which enables the business to more precisely focus retention initiatives. High-risk churners, for instance, can get unique offers to entice them to stay. The telecommunications business might execute particular retention strategies with the help of attrition projections and client groups. To decrease turnover and increase customer satisfaction, these techniques could include providing discounts, personalized offers, better customer service, loyalty programs, or focused marketing initiatives. The performance and accuracy of the model are constantly checked to guarantee its longevity. In order to adjust the model to changing client behavior and trends, new data or updated machine learning algorithms may be used.

**System Design**

Application Design (2.1)

The application design of the churn prediction system involves creating a user-friendly interface for data input and obtaining churn predictions. It includes the layout, user interactions, and integration with the machine learning models. The design ensures seamless user experience while capturing essential customer information for analysis. (Optional)

Process Flow (2.2)

The process flow illustrates how data and operations move through the system. In this context, it represents the steps taken to predict customer churn. The process includes data preprocessing, feature selection, model training (Logistic Regression, SVM, and Random Forest), evaluation, and generating churn predictions for each customer.

Information Flow (2.3)

Information flow outlines how data is passed between different components of the system. It shows how data is collected, transformed, and utilized throughout the churn prediction process. This includes the flow of customer data from Kaggle dataset columns such as customerID, gender, SeniorCitizen, etc., to the machine learning models for predictions.

Components Design (2.4)

Components design involves breaking down the system into its constituent parts. In this project, it includes designing the data preprocessing module, feature selection module, three separate machine learning models for Logistic Regression, SVM, and Random Forest, as well as result presentation components. Each component plays a crucial role in the overall churn prediction system.

Key Design Considerations (2.5)

Key design considerations encompass important factors that influence the system's architecture and functionality. For this project, considerations include data privacy, scalability, model interpretability, and real-time prediction capabilities. These aspects ensure the system meets business and ethical requirements.

API Catalogue (2.6)

The API catalogue lists and describes the APIs (Application Programming Interfaces) that allow external systems to interact with and access functionalities of the churn prediction system. It may include APIs for data input, model training, prediction retrieval, and result visualization, facilitating integration with other applications or services.

**Data Design**

8.1. Data Model:

For predicting customer churn, the data model needs to incorporate the provided parameters as features. This can be achieved by structuring the data in a tabular format, where each row corresponds to a customer and each column corresponds to a parameter such as gender, tenure, contract type, etc. Machine learning algorithms, such as logistic regression, decision trees, or neural networks, can then be trained using historical data to learn patterns and relationships between these parameters and customer churn. The data model should be designed to handle categorical variables (e.g., Partner, Internet Service) and continuous variables (e.g., Monthly Charges, Total Charges) appropriately to ensure accurate predictions.

8.2. Data Access Mechanism:

The data access mechanism involves collecting and managing the customer data parameters from various sources within the telecommunication company's systems. This might include customer databases, billing systems, service usage logs, and payment records. Integration tools or ETL (Extract, Transform, Load) processes can be used to gather and transform data from these sources into a format suitable for analysis. The data should be stored in a secure and accessible manner, potentially utilizing a database management system that supports querying and retrieval of the required parameters efficiently.

8.3. Data Retention Policies:

Data retention policies in this context refer to how long the telecommunication company will store the customer data parameters. Retaining historical data is crucial for training and testing churn prediction models accurately. The policies should specify the duration for which data will be kept before being archived or deleted. It's important to balance the need for retaining enough historical data for meaningful analysis with any legal or compliance requirements that govern data retention periods.

8.4. Data Migration:

Data migration might come into play when the telecommunication company updates its data infrastructure or switches to a different platform for data analysis. During data migration, the company should ensure that the integrity of the customer data is maintained. This includes mapping the old data schema to the new schema, handling any changes in data types or formats, and validating the accuracy of migrated data. Effective data migration is crucial to ensure that churn prediction models continue to perform reliably on the new platform.

**Interfaces**

**Data Collection and Integration Interfaces:**

* Data Sources: Interfaces to interact with various data sources within the telecommunication company, such as databases, logs, spreadsheets, etc. These interfaces ensure data extraction and collection for analysis.

**Data Preparation and Analysis Interfaces:**

* Data Preprocessing Scripts: Interfaces in the form of scripts (Python, R, etc.) that perform data preprocessing tasks like cleaning, handling missing values, and encoding categorical variables.
* Exploratory Data Analysis (EDA) Tools: Tools or code snippets that allow analysts to explore and visualize data patterns, correlations, and distributions.
* Feature Engineering: Interfaces for creating new features or transforming existing ones to enhance the predictive power of the model.
* Model Training Interfaces: Interfaces where machine learning models are trained using prepared data, involving algorithms like logistic regression, decision trees, etc.
* Hyperparameter Tuning: Interfaces to experiment with different hyperparameter settings to optimize the model's performance.

**Model Evaluation and Deployment Interfaces:**

* Model Evaluation Scripts: Interfaces for evaluating the trained model's performance using metrics such as accuracy, precision, recall, F1-score, etc.
* Model Interpretation Tools: Interfaces to interpret the model's predictions and understand the impact of different features on the prediction.
* Model Export: Interfaces to save the trained model to a file for later deployment.
* Deployment Interface: While you mentioned no UI, there might still be a basic deployment interface to load the trained model and make predictions on new data without a graphical UI.

**State and Session Management**

1.Data State: Keeping track of the current state of the data, such as its quality, preprocessing status, and any transformations applied. This helps ensure the consistency and accuracy of data used in model training and evaluation.

2. Model State: Monitoring the state of trained machine learning models, including their versions, hyperparameters, and performance metrics. This is crucial for tracking the evolution of models and understanding which configurations produce the best results.

3. Experiment Tracking: Maintaining a record of different experiments performed during the project, including variations in feature engineering, model selection, and hyperparameter tuning. This enables reproducibility and facilitates learning from past experiments.

4. Analysis Sessions: Considering each analysis or model training task as a "session." For instance, when conducting exploratory data analysis, feature engineering, or model training, you can define and manage these as separate analysis sessions.

5. Version Control: Viewing different versions of data preprocessing steps, model training, and evaluation as "sessions." Version control tools like Git can be used to manage changes across different phases of the project.

6. Logging and Documentation: Logging relevant information during different stages of the project serves as a form of session management. You can document details about data preprocessing steps, model configurations, evaluation results, and any important insights gained.

**Caching**

1. Data Preprocessing Caching:

* As you preprocess and clean the raw data, certain steps can be computationally intensive and time-consuming, such as handling missing values or encoding categorical variables.
* Caching the results of these preprocessing steps can save time during subsequent model training iterations or analysis sessions, as you won't need to repeat the same preprocessing steps each time.

1. Feature Engineering:

* Feature engineering involves creating new features from existing ones to enhance model performance.
* If certain feature engineering transformations are complex and time-consuming, caching the results can be beneficial. This avoids recomputation when experimenting with different model configurations.

1. Intermediate Results:

* During exploratory data analysis or model evaluation, you might calculate intermediate results or statistics that are used multiple times.
* Caching these intermediate results can prevent redundant calculations and speed up the overall analysis.

1. Model Training:

* If you experiment with different machine learning algorithms and hyperparameters, training models can be resource-intensive.
* Caching trained models or the results of training sessions can save time when comparing different model performances.

1. Experiment Tracking:

* As you experiment with various preprocessing techniques, model architectures, and hyperparameters, caching metadata about each experiment can help you organize and track your progress.

1. Data Exploration:

* While there might not be an interactive UI, caching visualizations or summary statistics from data exploration can be useful to revisit insights without recalculating everything.

**Non-Functional Requirements**

**Security Aspects:**

* Data Privacy: Ensure that sensitive customer information is handled and stored securely. Use encryption techniques to protect data at rest and during transit. Implement proper access controls to restrict unauthorized access to sensitive data.
* Secure Data Handling: Use secure coding practices to prevent vulnerabilities such as SQL injection or other attacks. Avoid exposing sensitive information in logs, error messages, or other outputs.
* Model Fairness and Bias: Ensure that the data used for training models is representative and unbiased. Detect and mitigate any potential bias in predictions to prevent unfair treatment of certain groups.
* Version Control: Keep track of changes in code, data preprocessing steps, and model configurations using version control systems to ensure traceability and accountability.
* Authentication and Authorization: Implement strong authentication mechanisms for any access to data, code, or model files. Use role-based access control to grant different levels of access based on user roles.

**Performance Aspects:**

* Data Processing Efficiency: Optimize data preprocessing steps to reduce computation time. Use efficient libraries and techniques for tasks like encoding categorical variables, handling missing data, and scaling features.
* Feature Selection: Identify the most relevant features to avoid unnecessary complexity and enhance model performance. This can also improve training efficiency.
* Model Selection and Hyperparameter Tuning: Experiment with different algorithms and hyperparameters to find the right balance between model complexity and predictive power.
* Parallel Processing: Utilize parallel processing techniques when applicable to speed up computations, especially for resource-intensive tasks like hyperparameter tuning.
* Model Optimization: Optimize model training and prediction processes for faster execution. Techniques like gradient boosting and optimized matrix operations can enhance model training speed.
* Scalability: Design your analysis pipeline and models to scale with growing datasets. Consider using distributed computing frameworks if dealing with big data.

**References:**

1. Iris Figalist, Christoph Elsner, Jan Bosch, Helena Holmstrom Olsson. “Customer Churn Prediction in B2B Contexts” [2019]
2. Pallav Routh, Arkajyoti Roy, Jeff Meyer. “Estimating customer churn under competing risks”. August 2020 Journal of the Operational Research Society.

1. Ahmad Hammoudeh, Malak Fraihat, Mahmoud Almomani. Selective Ensemble Model for Telecom Churn Prediction.[2019]
2. Debjyoti Das Adhikary, Deepak Gupta. “Applying over 100 classifiers for churn prediction in telecom companies.” Multimedia Tools and Applications[2020]
3. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019.
4. Ahmet Tuğrul Bayrak, Asmin Alev Aktaş, Orkun Susuz, Okan Tunalı. ”Churn Prediction with Sequential Data Using Long Short Term Memory”[2021]