Analytics Startup Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

Project	*Telecom customer Churn*
Requestor	*Prof. David Parent- Centennial College*
Date of Request	*July 14, 2025*
Target Quarter for Delivery	*Q3, August 13th 2025.*
Epic Link(s)	
Business Impact	*This project empowers telecom firms to shift from reactive to proactive customer retention and be able to accurately identify which key factors affect customer churn.* *By accurately identifying customers at risk of churning before they leave, the business can take early action—offering personalized offers, promotions, resolving dissatisfaction, and preserving long-term revenue. This not only prevents avoidable losses but also strengthens customer loyalty, optimizes marketing spend, and supports a smarter, data-driven retention strategy.*

1.0 Business Opportunity Brief

Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment.

……

The specific ask:

Clearly articulate the specific task you will be conducting to help achieve the opportunity

Business Context:

In today's highly competitive telecommunications market, customer retention is more cost-effective than customer acquisition. Telecom companies often face high churn rates, where customers discontinue their service and move to competitors, usually without warning. This directly impacts revenue, customer lifetime value (CLTV), and market share.

Business Opportunity / Ask:

This project aims to identify key drivers of customer churn and build a predictive model that helps the company proactively target at-risk customers. By leveraging internal data on customers, the business can uncover insights that inform strategic retention efforts.

To help reduce customer churn, this project will involve consolidating and analyzing customer data from multiple sources to uncover key drivers of churn and build predictive models. Using machine learning techniques such as logistic regression, decision trees, and XGBoost, I will estimate churn risk at the individual customer level. The analysis will include feature

engineering, model evaluation, and interpretation using tools like SHAP to ensure explainability. Finally, I will segment customers by churn risk and value, and provide targeted, data-driven recommendations to support proactive retention strategies and improve overall business performance.

1.1 Supporting Insights

Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally?

Canada's telecommunications industry is currently facing heightened competitive pressure due to an aggressive price war. Major incumbents such Rogers, Bell, and Telus—have been forced to lower their wireless plan prices in response to Quebecor's national expansion through Freedom Mobile. This has resulted in a 26.6% year-over-year drop in cellular service prices and a nearly 50% decline since 2019, benefiting consumers but placing significant strain on carrier profit margins. Earnings are expected to remain under pressure as more customers migrate to legacy plans with reduced pricing (Telecom Canada, 2024).

Customer retention has become a critical focus in this environment, with Rogers' postpaid churn rate falling slightly to 1.53% in Q4 2024, down from 1.67% in the previous year. Bell and Telus have reported similar churn performance, reinforcing the idea that low switching costs and commoditized service offerings continue to drive volatility in subscriber loyalty (Telecom Monitor, 2024).

The market share landscape remains highly concentrated, with Bell Group, Telus Group, and Rogers Group collectively controlling approximately 85.6% of Canada's telecom service revenue as of 2023. In terms of subscriber count, Rogers Wireless leads with 13.7 million customers, followed by Bell Mobility at 10.3 million and Telus Mobility at 9.5 million, representing a combined share of around 86% of the national wireless market (CRTC, 2024).

Key Competitors in the Canadian Telecom Market

The Canadian telecom industry is dominated by three national giants—Rogers Communications, Bell Canada, and Telus Communications—each offering full-service connectivity solutions. Alongside them, regional challengers such as Quebecor's Freedom Mobile, Eastlink, and the now-acquired Shaw Communications contribute to growing market competition.

Rogers Communications positions itself as a leader in digital connectivity, promoting the message of "Connecting Canadians with the latest in 5G and digital lifestyle experiences." Rogers operates a wide range of services including wireless (via Rogers Wireless, Fido, and Chatr), internet, cable TV, home phone, and a strong portfolio of media and sports properties such as Sportsnet and Citytv.

Bell Canada markets itself on network reliability and service breadth, branding its offering as "Canada's most reliable network with integrated communication and entertainment solutions." Bell provides mobile services through Bell Mobility, Virgin Plus, and Lucky Mobile, as well as internet (Fiber), satellite TV, home phone, and comprehensive enterprise communications.

Telus Communications distinguishes itself with a customer-first and socially conscious brand, emphasizing innovation and digital health under the message "Caring for customers and communities through technology." In addition to its mobile brands (Telus Mobility, Koodo, Public Mobile), Telus delivers high-speed internet, SmartHome Security, business tech solutions, and leads in the health-tech space through TELUS Health.

Quebecor, through Freedom Mobile, plays the role of the price disruptor in the market. Its value proposition centers on affordability—offering mobile plans up to 20% cheaper than national carriers. Operating primarily in Ontario, British Columbia, Alberta, and Manitoba, Freedom Mobile also offers home internet and IPTV bundles under the Vidéotron brand in Quebec.

1.2 Project Gains

Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI's.

Reduced Customer Churn:

This project empowers telecom firms to transition from reactive to proactive customer retention. By using churn prediction models, businesses can identify at-risk customers early and act swiftly with personalized interventions such as tailored promotions and offers ultimately improving retention and customer satisfaction.

Revenue Gains:

Predicting churn allows firms to retain more customers, especially high-value segments, thereby protecting revenue and reducing acquisition costs for new customers. Proactive intervention is significantly cheaper than reacquisition. Retaining an existing customer typically costs less than acquiring a new one.

Cost and Time Savings

Targeted promotional strategies and are more cost-effective than mass campaigns. Automated insights will also cut down on manual segmentation and guesswork, streamlining campaign planning. Marketing resources can be redirected from low risk to high-risk customers, improving ROI on campaigns and reducing irrelevant promotional offerings messaging

Reduced Wasted Marketing Spend:

Marketing resources can be redirected from low risk to high-risk customers, improving ROI on campaigns and reducing irrelevant promotional offerings messaging.

Quality Improvements:

Data-driven strategies improve the precision and relevance of retention efforts. Churn

interventions based on model predictions will be timely, targeted, and backed by measurable insights.

What Will Be Done Differently:

Unlike one-size-fits-all strategies, this project will use machine learning to analyze service, satisfaction, and demographic data at the individual level. Interpretable ML tools (like SHAP and LIME) will also help explain churn drivers to non-technical business teams.

Why Customers Will Care

More Relevant Communication: Customers will receive messages, offers, service adjustments and other incentives that align with their actual needs and usage, reducing frustration and enhancing satisfaction.

Faster Issue Resolution: By anticipating dissatisfaction early, customer support can act before a problem leads to churn, reinforcing trust and loyalty.

Implications of Doing Nothing

Continued Revenue Leakage: Without a predictive churn model, the business remains reactive, allowing preventable churn to persist, especially among high-revenue customers. Failing to act could mean continued revenue erosion, higher acquisition costs, misaligned marketing, and poor customer experience.

Misaligned Retention Strategies: Blanket discounts or generic loyalty campaigns can backfire, offering unnecessary incentives to loyal customers and ignoring those who are truly at risk of churning.

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.

2.0 Analytics Objective

List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.

Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed.

The objective of this analytics project is to predict customer churn and uncover the key factors driving it, enabling the telecom company to take proactive steps toward retention. By integrating demographic, service usage, satisfaction, financial, and geographic data, the project aims to build accurate and interpretable machine learning models that estimate each customer's likelihood to churn. The analysis will also segment customers based on churn risk and customer lifetime value (CLTV), supporting more targeted and cost-effective retention strategies. Ultimately, the goal is to deliver actionable insights that improve customer loyalty, reduce revenue loss, and support smarter, data-driven business decisions.

Key Business Questions

1. What are the key factors that drive customer churn?
2. Can we accurately predict which customers are at high risk of churning in the next quarter?

3. Which high-value customers are also at high churn risk and should be prioritized for retention?
4. What differences exist in churn behavior across regions, age groups, or service types?
5. What types of interventions, offers, and promotions are most likely to reduce churn among high-risk customer segments?

Assumptions

1. The provided datasets are representative of typical customer behavior and are consistent across quarters.
2. Churn is primarily impacted by current and recent interactions (e.g., last billing quarter, tenure to date, last offer accepted).
3. Customer satisfaction scores are a reliable proxy for sentiment and service quality perception
4. Customer churn is influenced by multiple dimensions—demographic, geographic, behavioral, and financial.
5. Missing data can be imputed with minimal bias; class imbalance can be addressed using techniques like SMOTE or class weighting.
6. External drivers of churn such as competitor offers or macroeconomic shifts are not explicitly captured.

Hypotheses to be Tested

1. Customers with shorter tenure are more likely to churn.
2. Churn is higher among month-to-month contract customers than among annual or two-year contract customers.
3. Low satisfaction scores (1–2 out of 5) significantly correlate with higher churn.

4. Customers without bundled or premium services (e.g., online security, streaming, tech support) are more likely to churn.
5. High CLTV customers who churn have identifiable churn reasons that could have been pre-emptively addressed.
6. Churn varies geographically, with some zip codes or cities having consistently higher churn rates.

2.1 Other related questions and Assumptions:

List any assumptions that may affect the analysis

1. Geographic Representativeness: Population data at the ZIP code level is assumed to be current and accurate, although some regions may be underrepresented.
2. Are churn rates higher among customers who have not accepted promotional offers or referred others?
3. Are certain combinations of services (e.g., streaming + security) associated with lower churn?

2.2 Success measures/metrics

What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.

Definition of Success

Success in this project will be defined by the ability to accurately predict customer churn, identify key churn drivers, and provide actionable insights that enable the business to reduce churn through targeted interventions. The outcome should improve retention, boost customer lifetime value, and align with company goals related to customer experience and revenue protection.

Category	KPI	Success Indicator
	Accuracy	≥ 80% accuracy on test data
	Precision, Recall, F1-score	Balanced precision and recall; F1 Score ≥ 0.70
	ROC	≥ 0.85 indicating strong discrimination between churners and non-churners
	Confusion matrix	Low false negatives (missed churners)

		prioritized over false positives
Business Impact	High-Risk Segment Detection	Identify top 15–20% of customers most likely to churn
		Empowers telecom firms to move from reactive to proactive customer retention.
		Revenue retention
		Firms will spend less on acquisition and more efficiently allocate retention offers
		helps the business focus its retention efforts where it matters most.
Customer Prioritization	CLTV	Segment customers into high-risk/high-value cohorts for action
Operational Readiness	Interpretability of Model	Key drivers are clear and explainable to business teams (feature importance ranked)

Drivers to Track During Analysis

- Customer Tenure
- Contract Type (Month-to-month vs. annual)
- Satisfaction Score
- Number of Services Used
- Payment Method
- Monthly and Total Charges
- Offer Type Accepted
- Referral Behavior
- Online Security, Streaming Services, Tech Support usage

2.3 Methodology and Approach

Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.

Exploratory Data Analysis (EDA), Logistic Regression, Random Forest, Decision Tree, XGBoost, Chi-square tests, SHAP/LIME explanations, SMOTE (for imbalance), and segmentation by churn & customer lifetime value (CLTV).

Type of Analysis:

- Classification Models:
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - XGBoost
    - Gradient Boosting
    - K-Nearest Neighbors (KNN)
    - Clustering K-Means or Hierarchical Clustering (grouping similar customers by risk profile or service usage)
    - SHAP
    - Visualization: Distribution plots, boxplots, feature importance charts, churn heatmaps
    - Feature Importance (from tree models)

Initial Approach:

The project will begin by identifying customers from the latest quarter in the Status dataset and defining the binary response variable:

- Churned = 1 (if Churn Label = Yes)

- Stayed = 0 (if Churn Label = No)

I will join this churn label with each customer's demographic, geographic, and service-related features from the remaining datasets. A Random Forest classifier will be the first exploratory model used to identify which variables (e.g., contract type, tenure, payment method, satisfaction score) are most predictive of churn.

I will then refine the analysis using logistic regression for interpretability and XGBoost to improve prediction performance. I will also apply Chi-square tests to assess statistical relationships between churn and categorical variables like Offer Type, Internet Service, and Payment Method.

Methodology:

1. **Data Integration & Preparation**

- Import and review all five datasets: status, services, demographics, location, and population.
- Merge them on the shared Customer_ID key to form a unified master dataset.
- Validate data consistency across merged fields.
- Handle missing values through appropriate imputation (e.g., median for numerical, mode for categorical), or exclusion if missingness is excessive.
- Identify and remove duplicate records.
- Standardize column names and normalize numerical variables (e.g., Monthly Charges, CLTV Proxy).
- Encode categorical variables using techniques such as Label Encoding or One-Hot Encoding, depending on model needs.

2. Exploratory Data Analysis (EDA)

- Visualize churn distribution to understand class imbalance.

- Perform univariate analysis (e.g., churn vs. satisfaction, tenure, contract type).
- Conduct bivariate and multivariate analysis using:
- Crosstabs and bar charts for categorical variables (e.g., churn vs. contract)
- Boxplots for numeric distributions by churn status
- Correlation matrix and heatmaps for numeric variables
- Conduct statistical hypothesis testing:
- Chi-square tests for independence (e.g., services vs. churn)
- T-tests or ANOVA for numerical features (e.g., tenure by churn)
- Identify outliers that may impact model performance.

3. Feature Engineering

- Create binary flags for all key services (e.g., Has_TechSupport, Has_Streaming, Has_OnlineSecurity).
- Construct derived features such as:
- Total_Services_Used
- Satisfaction_Category (e.g., low/medium/high based on rating)
- Contract_Length bucket
- CLTV_Proxy = Monthly Charges × Tenure
- Create interaction terms (e.g., Satisfaction × Services Used) to capture nonlinear effects.
- Bin population size into categories (e.g., rural, suburban, urban).

4. Addressing Class Imbalance

- Analyze churn class distribution.
- Use SMOTE (Synthetic Minority Oversampling Technique) or class weighting during model training to address imbalance.

- Validate that performance metrics (e.g., precision, recall) are not biased toward the majority class.

5. Model Development

- Split the dataset into training and test sets (e.g., 80/20 split with stratification).
- Train multiple supervised classification models:
- **Logistic Regression**: I'm using logistic regression as my baseline model because it's well-suited for binary classification problems like churn prediction. One of the key advantages is that it's easy to interpret — I can clearly see how each feature affects the probability of churn through its coefficients. This helps me establish an early understanding of which variables are most impactful before moving on to more complex models.
- **Decision Tree Classifier**: I included a decision tree model because of its intuitive structure. It allows me to build rule-based segments and see how the dataset is split into churn and non-churn groups based on key variables. I chose this model to help visually communicate findings to stakeholders who may not have technical backgrounds, as it provides simple, understandable if-then rules.
- **Random Forest Classifier**: I selected the random forest because it builds on the decision tree by combining multiple trees and averaging their outputs. This makes the model more robust and reduces the risk of overfitting. It also helps capture interactions between variables that a single tree might miss. I'm using it to improve predictive accuracy while still keeping a level of interpretability through feature importance.
- **XGBoost Classifier**: I chose XGBoost because it's known for delivering high performance on structured datasets. It supports regularization, handles missing values effectively, and is often used in winning solutions for classification problems. In my

project, I'm using it to push the limits of model performance and explore the most accurate prediction of churn risk.

- **Gradient Boosting**: I'm including Gradient Boosting in my modeling process to improve prediction accuracy by capturing complex, non-linear relationships in the data. It's particularly well-suited for structured churn datasets and allows for extensive tuning to balance performance with overfitting. This model will serve as one of my strongest classifiers alongside Random Forest and XGBoost, and its results will be interpreted using SHAP values for transparency.
- Apply cross-validation and hyperparameter tuning (e.g., GridSearchCV) for each model.
- Evaluate using key classification metrics: Precision, Recall, F1-Score, ROC-AUC.

6. Model Explainability

- Apply SHAP (Shapley Additive Explanations) to interpret global and individual predictions.
- Use LIME (Local Interpretable Model-Agnostic Explanations) to explain model decisions on select customers.
- Generate visual plots showing top features influencing churn risk (e.g., satisfaction, tenure, contract type).

7. Churn Segmentation & CLTV Mapping

- Use the CLTV_Proxy and model's churn probabilities to segment customers into four quadrants:
- High Churn / High Value
- High Churn / Low Value
- Low Churn / High Value
- Low Churn / Low Value

- Define tailored retention strategies for each segment:
- High-value churners: loyalty offers or proactive outreach
- Low-value churners: cost-effective retention or exit

8. Output & Strategic Recommendations

- Visualize churn insights through charts, bar graphs, and correlation heatmaps.
- Present an executive dashboard or slide deck with:
- Key findings from the models
- Region- or segment-specific churn patterns
- Churn drivers for key customer profiles
- Provide business recommendations:
- Which customer segments to prioritize
- Recommended retention actions (e.g., contract upgrades, bundled offers)
- Messaging strategies per churn segment

3.0 Population, Variable Selection, considerations

Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today.

Audience/population selection: 7043 customers

Observation window: California, 3rd Quarter, 2019

Inclusions: Customers with complete status and service profiles. All demographic and location attributes with valid entries.

Exclusions: Records missing churn labels or key demographic variables, and Duplicate customer IDs.

Data Sources: https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113/data

Audience Level: Management

Variable Selection:

Dimension	Sample Variables
Demographics	Age, Gender, Marital Status, Senior Citizen, Dependents
Services	Contract Type, Tenure, Offer Accepted, Internet Service, Charges
Satisfaction	Satisfaction Score, Churn Score, Churn Category/Reason
Financial	CLTV, Monthly Charges, Total Charges, Total Refunds
Geography	State, City, Zip Code, Population

Derived Variables:

Assumptions and data limitations:

- No direct revenue per customer; must proxy CLTV
- Imbalanced target variable
- Unknown external competitor effects not captured in the data
- Data Completeness, Geographic Bias
- (Some ZIP codes may be underrepresented or overrepresented.)

4.0 Dependencies and Risks

Identification of key factors that may influence the outcome of the project and likelihood of it happening:

Risk	Likelihood (based on historical data)	Delay (based on historical data)	Impact
Churn rate being inflated by counting multiple contracts from the same rooftop as individual observations.	*Low*		*Once analysis begins, we can quantify the inflation. However, this approach allows us to compare how the same dealer performed across different contracts and find useful patterns.*

Risk	Likelihood	Delay Risk	Impact / Mitigation
Incomplete or inconsistent customer IDs across datasets (e.g., services, status, demographics)	Medium	Moderate	Could prevent accurate merging. Will use inner joins and row tracking to quantify mismatches; investigate and impute IDs.

Risk	Likelihood	Delay Risk	Impact / Mitigation
Pre-modeled churn score columns (e.g., Churn Score) introduce data leakage	High	Low	These features will be **excluded** from model training but used for validation/benchmarking only.
Skewed churn rate (class imbalance) leads to biased prediction models	High	Low	Will use **stratified sampling**, and apply **SMOTE or weighted classifiers** to balance predictions.
Duplicate records for same customer in different quarters or files	Medium	Low	Duplicates will be identified and dropped based on CustomerID and time context.
Outliers in financial data (e.g., unusually high Total Charges or Refunds)	Medium	Low	Will conduct outlier detection and decide whether to cap or exclude those records during preprocessing.
Misinterpretation of service bundle columns (e.g., multiple "Yes"/"No" flags)	Low	None	Cross-validation will be used to ensure accurate variable encoding and avoid redundancy in model input.
Geographical bias due to overrepresented zip codes	Medium	Low	Normalize location-level data using population file. Consider aggregating for regional insights.

Risk	Likelihood	Delay Risk	Impact / Mitigation
Model interpretability concerns for black-box algorithms (e.g., XGBoost)	Low	None	Use SHAP and feature importance to explain high-performing models in stakeholder-friendly language.
Delay in stakeholder feedback or pilot decision after model presentation	Medium	High	Build interim check-ins into project plan (e.g., Story Board 1 + Go/No-Go) to avoid last-minute surprises.

5.0 Deliverable Timelines

List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process.

Grade Modules	**Due Date**
BA706 Revision	Monday July 14
Analysis Plan & Data Finalization	Monday July 14
Data Exploration	Monday July 21
Peer Review Week 4	Monday July 28
Modeling	Monday August 4
Governance	Wednesday August 13
Documentation	Wednesday August 13
Peer Review Week 6	Wednesday August 13
Presentation	August 11-13
Portfolio	*Wednesday August 13*

Bibliography

Government of Canada, Canadian Radio-television and Telecommunications Commission (CRTC). (2025, March 7). *Canadian Telecommunications Market Report 2025*. CRTC. https://crtc.gc.ca/eng/publications/reports/policymonitoring/2025/ctmr.htm

Government of Canada, Canadian Radio-television and Telecommunications Commission (CRTC). (2024, October 7). *Competition in the wireless services market*. CRTC. https://crtc.gc.ca/eng/phone/mobile/indus.htm

IQMetrix. (2025, February 28). iQmetrix. https://www.iqmetrix.com/blog/canadian-tier-1-carrier-report-highlights-increased-competition-heightens-market-intensity

Telco customer churn (11.1.3+). (2019, November 8). Kaggle. https://www.kaggle.com/datasets/ylchang/telco-customer-churn-1113/data