Turning Churn into Opportunity

David S McMillan Jr.

Western Governors University

Table of Contents

A. Project Highlights 4
B. Project Execution 5
C. Data Collection Process 6
C.1 Advantages and Limitations of Data Set 6
D. Data Extraction and Preparation 6
E. Data Analysis Process 7
E.1 Data Analysis Methods 6
E.2 Advantages and Limitations of Tools and Techniques 7
E.3 Application of Analytical Methods 7
F Data Analysis Results 8
F.1 Statistical Significance 7
F.2 Practical Significance 8
F.3 Overall Success 8
G. Conclusion 11
G.1 Summary of Conclusions 9
G.2 Effective Storytelling 9
G.3 Recommended Courses of Action 9
H Panopto Presentation 9
References 14

A. Project Highlights

In my capstone project, I ask: Does contract length affect customer churn rates in the telecommunications industry? Specifically, the project investigated whether contract length (month-to-month, one-year, or two-year agreements) influences customer churn rates in the telecommunications industry. The business need is to understand what drives customer churn so they can develop effective retention strategies. The project scope included statistical hypothesis testing using the chi-square test to examine the relationship between contract type and churn status, along with exploratory data analysis and data visualization to communicate findings effectively. The scope explicitly excluded machine learning techniques, such as predictive modeling and customer segmentation via clustering, due to time constraints and project requirements.

The solution employed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to structure the analysis workflow, progressing systematically from business understanding through data preparation, statistical modeling, evaluation, and deployment of findings. The project utilized Python 3.9 as the primary programming environment, leveraging pandas for data manipulation, scipy for statistical testing, and matplotlib and seaborn for data visualization. Analysis was conducted in Jupyter Notebook to enable interactive code development and on the fly data manipulation. The analytical approach focused performing chi-square hypothesis testing to determine statistical significance at $\alpha = 0.05$, calculating Cramér's V to measure effect size, and creating professional visualizations, including a bar chart comparing churn rates by contract type and a heatmap displaying the contingency table structure. This approach provides clear and statistically sound evidence about the relationship between contract length and customer churn.

B. Project Execution

The goal of this project was to determine whether contract type significantly impacts customer churn rates in the telecommunications industry. The project consisted of downloading the Telco Customer Churn dataset and importing it into Jupyter Notebook for statistical analysis. The project provides clear statistical evidence, using chi-square hypothesis testing, to demonstrate if the relationship between contract length and Churn is significant enough to inform business strategy. The final deliverables included a comprehensive written analysis, two professional visualizations showing churn patterns across contract types, and actionable recommendations for the telecommunications company.

The dataset used was the IBM Telco Customer Churn dataset, which I downloaded as a CSV file from Kaggle. Upon downloading, I used Jupyter Notebook with Python to load and explore the dataset. I used the Pandas library to import the CSV file, NumPy for numerical operations, SciPy for the chi-square test of independence, and the Matplotlib and Seaborn libraries to create two visualizations: a bar chart showing churn rates by contract type and a heatmap of the contingency table. In performing this analysis, I provided statistical evidence that contract type significantly affects churn rates, giving our telecommunications company the information needed to make data-driven decisions on contract strategy and retention initiatives.

The methodology used for this project was CRISP-DM (Cross-Industry Standard Process for Data Mining). I chose CRISP-DM because of its structured approach specifically designed for data analytics projects. The methodology guided the project through six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This framework ensured systematic progress from defining the research question through delivering final recommendations. The process was straightforward because the dataset was already clean and the research question was focused exclusively on testing one specific relationship using chi-square analysis. Listed below are the steps I followed:

- Business Understanding: Defined research question and hypothesis

- Data Understanding: Loaded the dataset and explored the Contract and Churn variables
- Data Preparation: Created a simplified dataframe with only the necessary columns
- Modeling: Performed a chi-square test of independence and calculated the effect size
- Evaluation: Interpreted statistical significance and practical significance
- Deployment: Created visualizations and documented findings with business recommendations

I completed the project timeline as planned over 9 days from January 13-21, 2026. The only variance from the original timeline was that data cleaning required less time than anticipated because the dataset had minimal quality issues. I completed all milestones on schedule, and I used the time saved during data preparation to enhance the overall quality of the report and presentation.

C. Data Collection Process

The data selection and data collection plan did not differ from my plan in Task 2. The plan was to use the IBM Telco Customer Churn dataset to answer my research question about whether contract type influences churn rates, which was publicly available on Kaggle.com. The dataset was downloaded as a CSV file and loaded directly into Jupyter Notebook using the pandas read_csv function.

I did not encounter any obstacles during the collection process, as the dataset was freely accessible and could be downloaded as a CSV file. This allowed me to easily execute the plan outlined in Task 2. The download process was straightforward, requiring only a free Kaggle account to access the dataset. Once downloaded, the CSV file was placed in the same directory as my Jupyter Notebook, making it immediately available for analysis without any file path or access issues.

I did not encounter any unplanned data governance issues for this process, as it was planned based on the work already completed in Task 2. The dataset contains no personally identifiable information, and the dataset is unrestricted for educational and research use under Kaggle's terms of service. All data governance considerations—including privacy, security, and ethical use were handled exactly as outlined in Task 2. I needed no additional precautions or adjustments during the project execution.

C.1 ADVANTAGES AND LIMITATIONS OF DATA SET

Below are a couple advantages and disadvantages of working with this dataset:

Advantages:

- The dataset was already prepped and clean, which made it easy to work with. For example, there were no duplicate records, and minimal missing values. Having a high-

quality data set required little time to clean the data, allowing me to focus more time on the statistical analysis itself.
- The dataset had all the necessary fields required to perform the tasks outlined to complete this work. For example, it included both variables needed for chi-square testing: Contract type (with three clear categories: month-to-month, one-year, two-year) and Churn status (binary Yes/No). This relationship between the dataset structure and the research question eliminated any need for complex data transformation.

Disadvantages:

- The dataset is limited to a single point in time, representing a cross-sectional snapshot of customer status rather than being yearly/seasonal. For example, having a dataset that tracked customers over multiple years would be beneficial to evaluate whether the relationship between contract type and churn changes as market conditions evolve. The lack of temporal data means I cannot assess churn trends over time.
- The dataset comes from a single telecommunications company, which limits the breath of findings. For example, results may be specific to this company's customer base, pricing structure, or service quality rather than representing universal patterns across the entire telecommunications industry. Different telecom companies, having different markets or regions may produce different relationships between contract type and churn.

D. Data Extraction and Preparation

The process of data extraction for this project was straightforward. I became a Kaggle member first, then I downloaded the CSV file from Kaggle.com. Once the dataset was downloaded to my computer, I imported it into my Jupyter Notebook using Pandas' read_csv function. Once I imported the dataset into a dataframe, I moved on to preparing the data. In preparation, I reviewed the basic structure and gained insight into the dataframe. I specifically examined the dataset's shape, the data types of each column, remove duplicates and identified null values using the isnull() function. In preparation for the analysis portion of this project, I created a simplified dataframe containing only the two columns needed for the chi-square test: Contract and Churn. This extraction approach was appropriate because my research question focused exclusively on the relationship between these two variables. I created a focused dataframe and was ready to proceed with the statistical analysis.

E. Data Analysis Process

E.1 DATA ANALYSIS METHODS

Statistical hypothesis testing was the primary method used to complete this project. This method was necessary to answer the research question about whether contract type significantly impacts customer churn rates. The hypothesis was that customers on month-to-month contracts would

have dramatically higher churn rates than those on one- or two-year contracts. I performed the chi-square test, which allowed me to examine the relationship between contract type. I found this to be the most appropriate method because the chi-square test is used to determine whether two variables are associated. The test produces a p-value and an effect size measure (Cramér's V), which help quantify the strength of the relationship. I also created visualizations—a bar chart and a heatmap—to present the findings.

E.2 ADVANTAGES AND LIMITATIONS OF TOOLS AND TECHNIQUES

The tools I used to complete this project and produce the data analytical solution were Python and Jupyter Notebook. I used the following Python libraries: Pandas, NumPy, SciPy, Matplotlib, and Seaborn. I used the Pandas library to import the dataset into my Jupyter Notebook, which enabled me to complete all the work for this project, including data manipulation and creating the contingency table. The NumPy library was used to perform mathematical functions, such as calculating the square root, when computing Cramér's V. The SciPy library was used to perform the chi-square test, which provided the test statistic, p-value, degrees of freedom, and expected frequencies. I used the Matplotlib and Seaborn libraries to create the bar and heatmap visualizations.

Advantages:

- Using Jupyter Notebook allowed me to have an easy-to-manage environment for importing data, cleaning it, performing the statistical analysis and documenting findings in a simple, clean workspace.
- Python's data visualization libraries (Matplotlib and Seaborn) allowed for complete customization of chart appearance, including colors, labels, annotations, and formatting.

Limitations:

- Jupyter Notebook is excellent for analysis and documentation while it is not designed for creating formal written reports. The final Task 3 report was written in Microsoft Word, requiring any statistics, and visualizations from the notebook to the document to be manually copy and pasted.
- Python's statistical testing requires understanding the underlying functions and their outputs. Unlike point-and-click statistical software, Python requires writing code to execute tests. Without a basis in Python code, this would prove to be a steep learning curve.

E.3 APPLICATION OF ANALYTICAL METHODS

The following analytical methods and steps were used:

1. Explored the dataset to understand the distribution of customers across contract types (month-to-month, one-year, two-year) and churn status (Yes, No) using value counts and percentages.
2. Created a contingency table using pandas crosstab function showing the count of customers in each Contract Type × Churn Status combination. This table formed the foundation for the chi-square test.
3. Verified that the dataset met chi-square test requirements: each customer appeared only once (independence assumption), both variables were categorical (Contract and Churn), and the sample size was sufficient for valid testing.
4. Calculated expected frequencies from the chi-square test output and verified that all expected cell frequencies were ≥ 5, confirming the chi-square test assumptions were met and results would be statistically valid.
5. Conducted chi-square test of independence using scipy.stats.chi2_contingency function, which automatically calculated the chi-square statistic, p-value, and degrees of freedom.
6. Calculated Cramér's V effect size to measure the strength of the relationship between contract type and churn, providing context beyond just statistical significance.
7. Compared the p-value (< 0.0000) to the alpha value (0.05) to determine whether the relationship between contract type and churn was statistically significant.
8. Visualized the churn rates by contract type using a bar chart to show the magnitude of differences (42.71% vs 11.27% vs 2.83%) and make findings accessible to stakeholders.
9. Created a heatmap of the contingency table to provide a visual representation of customer distribution across contract types and churn outcomes, supporting the statistical findings.
10. Verified that all visualizations accurately represented the data by cross-checking chart values against the original contingency table and calculated percentages to ensure no errors in data display.

F Data Analysis Results

F.1 STATISTICAL SIGNIFICANCE

To assess the statistical significance of my solution, I performed a chi-square test on two categorical variables: contract type and churn status. The chi-square statistic measures the discrepancy between observed and expected frequencies under the null hypothesis. A larger chi-square value would indicate a greater separation from independence. The significance level was set to 0.05, which was used to determine our statistical significance. a value less than 0.05 means we have statistical significance, while a value greater than 0.05 indicates there is a lack of or little to no statistical significance. The null hypothesis would be if the contract type and churn status are independent and there is no relationship between the two attributes.

In contrast, my alternative hypothesis states that contract type and churn status do have a relationship, with contract type significantly affecting churn rates. Here is the test I performed to either reject or fail to reject the null hypothesis. See below for the results of the chi-square test.

Chi-Square Test Results:

- Chi-square statistic (χ^2): 1184.60
- p-value: < 0.0000
- Degrees of freedom: 2
- Cramér's V (effect size): 0.41
- Significance level (α): 0.05

The test resulted in a chi-square statistic of 1184.60, a significant value indicating a substantial departure from independence. The p-value from the test was zero. This value is below the 0.05 significance level threshold, so we can easily conclude that there is statistical evidence of a relationship between Contract and Churn. A Cramér's V effect size of 0.41 indicates a moderate effect, indicating that the relationship is not only statistically significant but also meaningful. Based on the chi-square statistic and p-value, we have sufficient evidence to reject the null hypothesis and support our alternative hypothesis.

This is a significant analysis for the telecommunications dataset, as contract type is a factor a company can control. Month-to-month contracts had a churn rate of 42.71%, while one-year contracts had a churn rate of 11.27% and two-year contracts had a churn rate of 2.83%. This validation of the alternative hypothesis indicates that longer contract commitments dramatically reduce Churn, which a business enterprise can use to design retention strategies and incentive programs to encourage customers to switch to long-term agreements.

F.2 PRACTICAL SIGNIFICANCE

The practical significance of this project is to identify which contract types are most strongly associated with customer retention. Seeing that we proved there is significance in the relationship between Contract and Churn data, a telecommunications company can begin building better strategies to retain its customers and reduce Churn.

Based on the results of this analysis, month-to-month customers represent the highest churn risk segment, with churn rates more than 15 times higher than two-year contract customers. Reducing Churn in the month-to-month segment can prevent substantial revenue loss. The larger the difference in churn rates, the greater the opportunity for targeted intervention. Based on these results, a company could, for example, create a campaign to convert month-to-month customers to longer-term contracts that include bonus incentives for employees and discounts to customers.

F.3 OVERALL SUCCESS

Overall, I view this project as a success. The goal of the project was to determine whether contract type significantly influences customer churn rates and whether, with the help of statistical evidence telecommunications companies can make informed business decisions about retention strategy. Business recommendations provided: The analysis identified month-to-month customers as the highest-risk segment

I was able to provide statistical evidence with both significance testing and effect size measurement that has been outlined in F.1 and F.2. The chi-square test was executed adequately with valid results, statistical assumptions were verified, the hypothesis was clearly evaluated, and the visualizations provide further proof the relationship of the data in question. This analysis will help inform the most effective contract strategy and retention decisions, with the goal of reducing churn rates. The project demonstrates that contract length is a factor in customer retention.

G. Conclusion

G.1 SUMMARY OF CONCLUSIONS

The project's main objective was to determine whether contract type significantly impacts customer churn rates in the telecommunications dataset provided by IBM and to provide statistical evidence to inform retention strategy decisions. To do this, I gathered and explored the IBM Telco Customer Churn dataset using Python in Jupyter Notebook, with Pandas, NumPy, SciPy, Matplotlib, and Seaborn. I determined statistical significance using a chi-square test of independence between two categorical variables: contract type and churn status. The results of this test demonstrated a highly significant relationship with a chi-square statistic of 1184.60 and a p-value equal to zero. The effect size, measured by Cramér's V = 0.41, indicated a moderate effect, demonstrating that the relationship is not only statistically viable but also meaningful. This validates my hypothesis that contract type significantly affects churn rates and this would suggest that longer contracts will aid in reducing Churn.

Through exploratory analysis, I evaluated churn rates across three contract types: month-to-month (42.71%), one-year (11.27%), and two-year (2.83%). I created visualizations, including a bar chart showing churn rates by contract type and a heatmap of the contingency table distribution of customers across contract and churn categories. These results provide enough evidence that contract length is a critical factor in customer retention and that month-to-month customers are more at risk of departure.

G.2 EFFECTIVE STORYTELLING

The visualizations used in this project are a great match for the data as they provide a clear representation of the relationship between Contract and Churn. The visualizations for this project were created using the Matplotlib and Seaborn Python libraries. The bar chart shows the churn rates by contract type and provides a clear picture of the differences in churn behavior. This visualization was used to effectively and visually communicate the hypothesis findings. It clearly shows that month-to-month contracts have churn rates more than 15

times higher than those of two-year contracts. The red dashed line, on the bar graph, represents the overall average churn rate (26.54%). This establishes context for stakeholders to understand how each contract type compares to the baseline.

The contingency table heatmap supports the scope of this project by visually showing the distribution of customers across contract types and churn outcomes. The color-coded cells make it readily apparent that month-to-month contracts have a high churn rate. Also, it's apparent that month-to-month agreements have a high churn rate.

Both of these visualizations provide us with evidence that answers the research question of whether contract type influences churn rates. The combination of the bar chart (showing percentages) and the heatmap (showing raw counts) provides both statistical clarity and business context. These visualizations demonstrate that customers on longer-term contracts have dramatically lower churn rates.

G.3 RECOMMENDED COURSES OF ACTION

The analysis throughout this project aimed to answer the question: Does contract type significantly affect customer churn rates in the telecommunications industry? Based on this analysis, we found, through both statistical hypothesis testing and exploratory analysis, that contract length has a highly significant relationship with Churn, with month-to-month customers churning at rates more than 15 times higher than those of two-year contract customers. The chi-square test results and the differences in churn rates provide strong evidence that contract type is a critical lever for reducing Churn. These findings form the foundation for actionable business recommendations.

Recommendation 1: Implement Contract Conversion Incentive Programs

My first recommendation is to develop targeted incentive programs designed to convert month-to-month customers to longer-term contracts. Month-to-month customers represent the highest-risk segment with a churn rate of 42.71%/ That's more than three times the overall average and substantially higher than the 11.27% rate for one-year contracts. Even encouraging customers to commit for just one year rather than remaining month-to-month would dramatically reduce churn risk.

I recommend that if a company were to act on this analysis, it offer financial incentives such as a 7-12% discount on monthly charges for customers who switch from month-to-month to annual contracts. Committing customers to one-year contracts would substantially reduce churn-related revenue loss and would likely offset the cost of providing the discount. This recommendation directly addresses the research question by leveraging the proven relationship between contract length and churn to design a targeted churn-retention strategy.

Recommendation 2: Redesign Default Contract Offerings and Sales Approach

The second recommendation I am going to make focuses on preventing customers from selecting month-to-month contracts in the first place rather than converting them after the fact. I recommend that a company redesign its contract offerings and sales approach to make longer-term contracts the default. Currently, we have 55% of customers on a month-to-month contract, creating a huge, high-risk customer base. By restructuring initial offerings, we can shift more new customers toward annual commitments from the start.

Additionally, the sales process should emphasize the value and cost savings of yearly contracts as the recommended option, positioning month-to-month as a premium option rather than the standard choice. Based on the analysis showing that two-year customers churn at only 2.83%, the company should also introduce aggressive promotional offers for two-year commitments.

This approach addresses the research question by proactively managing contract mix rather than reactively responding to Churn after customers elected to have month-to-month preference. If proven successful, this strategy would greatly reduce the company's baseline churn rate. This recommendation is based on the finding that contract length is a significant factor in churn behavior. This makes contract selection during our initial customer acquisition a strategic lever for long-term retention success.

References

No sources needed or cited