CUSTOMER CHURN PREDICTION

A report on
Machine Learning Lab Project
[CSE-3183]

Submitted By

ANKIT KUMAR	**210962158**
MIHIR JATINKUMAR PATEL	**210962192**

MANIPAL
ACADEMY *of* HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY,
MANIPAL ACADEMY OF HIGHER EDUCATION

NOVEMBER 2023

Customer Churn Prediction

Ankit Kumar
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal Academy of Higher Education, India
ankitk7519@gmail.com

Mihir Patel
Department of Computer Science and Engineering
Manipal Institute of Technology
Manipal Academy of Higher Education, India
mihirp741@gmail.com

***Abstract*— This project presents a comprehensive machine learning approach to predict customer churn in a telecommunications company. Utilizing a rich dataset, we employ various data pre-processing techniques, including encoding of categorical variables, feature scaling, and handling missing values. The study explores multiple machine learning algorithms like Logistic Regression, Random Forest, and Support Vector Machines, assessing their performance based on accuracy, precision, recall, and F1 score. Key insights from exploratory data analysis and the model's feature importance are discussed. The project culminates in the deployment of a tuned Logistic Regression model, offering a user-friendly interface for real-time churn predictions. This work not only enhances customer retention strategies but also provides a framework for future research and applications in similar domains.**

Keywords— ***Customer Churn Prediction, Machine Learning, Logistic Regression, Predictive Analytics, Feature Engineering, Data Pre-processing, Model Evaluation***

I. Introduction

In the dynamic landscape of the telecommunications industry, customer retention emerges as a paramount challenge. This project addresses the critical issue of customer churn – a phenomenon where customers cease their relationship with a service provider. Churn not only impacts a company's revenue but also serves as a barometer for customer satisfaction and service quality.

This report delves into developing a predictive model to identify potential churners, leveraging machine learning techniques. By systematically analysing customer data, the model aims to predict churn likelihood, enabling proactive retention strategies. We explore various algorithms, assess model performance, and implement a solution that combines predictive accuracy with practical usability. This approach not only aids in decision-making processes but also contributes to the broader understanding of customer behaviour patterns within the telecom sector.

Our methodology encompasses data pre-processing, exploratory analysis, model selection, and tuning, culminating in a user-friendly interface for real-time predictions. This comprehensive approach underscores the significance of advanced analytics in customer relationship management and offers insights into the applicability of machine learning in business contexts.

II. Literature Review

1. Review of Churn Prediction in Telecom Industry [1]

This paper provides an in-depth review of various data mining techniques applied in the telecom industry for churn prediction. It highlights the uniqueness of telecom datasets and discusses how different machine learning models, including decision trees, neural networks, and support vector machines, are adapted for churn analysis. The paper also examines the importance of feature selection and the challenges posed by the imbalanced nature of churn datasets.

2. Ensemble Methods in Churn Prediction [2]

This review focuses on the application of ensemble methods like bagging and boosting in churn prediction. It emphasizes how these techniques improve prediction accuracy by addressing overfitting and variance in the model. The paper also compares the performance of ensemble methods with single predictive models, demonstrating the superiority of ensemble methods in handling complex churn prediction problems.

3. Customer Churn Prediction Using Advanced Machine Learning Techniques [3]

This comprehensive survey delves into various advanced machine learning approaches for churn prediction, including deep learning models. It discusses the evolution of machine learning techniques from traditional models to more complex structures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The paper also addresses the challenges faced in implementing these advanced models, such as data scarcity and computational complexity.

4. Customer Churn Prediction Using Advanced Machine Learning Techniques [4]

This literature review specifically addresses the issue of class imbalance in customer churn datasets. It explores various strategies to handle this imbalance, such as resampling techniques, cost-sensitive learning, and anomaly detection methods. The paper provides a critical analysis of each method's effectiveness and practicality in real-world scenarios.

5. Impact of Customer Satisfaction Data on Churn Prediction [5]

This review examines the impact of incorporating customer satisfaction metrics into churn prediction models. It emphasizes the predictive value of customer satisfaction data, often overlooked in traditional models that focus solely on usage patterns and demographic information. The paper demonstrates how integrating customer feedback can significantly enhance the accuracy of churn prediction models.

6. Application of SVM and Parameter Selection in Churn Prediction [6]
This review delves into the utilization of Support Vector Machines (SVM) in churn prediction, particularly in subscription-based services. It highlights the effectiveness of SVM in handling high-dimensional data and compares two different parameter-selection techniques: grid search and genetic algorithms. The study emphasizes the importance of choosing the right parameters in SVM to improve prediction accuracy and provides insights into the trade-offs between computational efficiency and model performance.

7. Improving Balanced Random Forests for Churn Prediction [7]
This paper reviews an enhanced version of the balanced random forests algorithm tailored for churn prediction. It discusses the limitations of traditional random forest in dealing with class imbalance, a common issue in churn datasets. The review elaborates on how the improved algorithm adjusts the class distribution in the training set and alters the decision threshold to better handle imbalanced data, leading to more accurate and reliable churn predictions.

8. Comparative Study of Multiple Classifiers in Churn Prediction [8]
While primarily focused on stock price prediction, this paper provides valuable insights for churn prediction by comparing multiple classifiers. It discusses various model evaluation metrics and the importance of choosing the right metric for model comparison. The study's methodology and findings can be directly applied to churn prediction, particularly in evaluating and selecting the most effective machine learning model for a given dataset.

9. Data Mining Techniques in Telecom Churn Management [9]
This literature review focuses on the application of various data mining techniques in managing customer churn in the telecom sector. It covers a range of techniques from simple logistic regression to complex neural networks and decision trees. The paper also discusses the role of data preprocessing and feature selection in improving model performance and provides case studies demonstrating the successful application of these techniques in real-world scenarios.

10. Churn Prediction in the Context of Customer Relationship Management [10]
This review offers a broader perspective, situating churn prediction within the context of customer relationship management (CRM). It explores how churn prediction models can be integrated into CRM systems to not only predict churn but also to provide actionable insights for customer retention strategies. The paper discusses the evolution of churn management techniques, from traditional statistical models to advanced machine learning algorithms, and forecasts future trends in the field.

III. Objectives

1. Develop an Efficient Feature Extraction Mechanism: Create a model that can efficiently extract relevant features from customer data, potentially using advanced deep learning techniques that balance performance and computational feasibility.

2. **Enhance Model Generalizability:** Design a model that performs well across various datasets and can be adapted to different industries beyond telecommunications, ensuring broader applicability.

3. **Incorporate Complex Customer Behavior Analysis:** Integrate advanced analytics to capture evolving customer behavior patterns and interactions with services, enhancing the predictive power of the model.

4. **Optimize Data Transformation Methods:** Investigate and apply effective data transformation and preprocessing strategies to improve model accuracy and robustness.

5. **Balance Complexity with Interpretability:** Aim for a model that not only provides high accuracy but also maintains a level of interpretability, allowing business users to understand and trust its predictions.

6. **Facilitate Real-time Prediction and System Integration:** Develop the model for real-time churn prediction and ensure it can be integrated seamlessly with existing CRM systems for practical use.

7. **Address Sample Imbalance and Overfitting:** Implement techniques like Synthetic Minority Over-sampling Technique (SMOTE) and appropriate validation methods to handle sample imbalance and reduce overfitting risks.

IV. Research gaps

Despite advancements in customer churn prediction (CCP) within the telecommunication industry, several research gaps remain:

1. **Inadequate Feature Extraction**: Many existing models struggle with efficient feature extraction, leading to suboptimal predictive performance. Deep learning models, while promising, often require substantial computational resources and expertise, which may not be feasible for all organizations.

2. **Limited Generalizability Across Diverse Data Sets:** Current models are often tailored to specific datasets and may not perform well when applied to data from different sources or industries. This limits their utility in varied business contexts.

3. **Overlooked Customer Behaviour Patterns**: Many models fail to capture complex customer behaviour patterns and interactions with various services. As customer preferences evolve rapidly, models need to adapt and incorporate these changes effectively.

4. **Challenges in Data Transformation and Pre-processing:** Effective data transformation and pre-processing strategies are crucial but often overlooked. The impact of

different data transformation methods on model performance is not fully explored.

5. **Balance Between Model Complexity and Interpretability:** While complex models like deep neural networks can offer higher accuracy, they often lack transparency and interpretability, making it difficult for business users to understand and trust the model's predictions.

6. **Real-time Prediction and Integration:** Many churn prediction models are not designed for real-time analysis and lack integration capabilities with existing customer relationship management (CRM) systems.

7. **Sample Imbalance and Overfitting Risks:** Models often struggle with imbalanced datasets where churned customers are a minority. This leads to overfitting and poor generalization to new data.

V. Methodology

The methodology for the Customer Churn Prediction project is comprehensive and detailed, encompassing various stages from data pre-processing to model evaluation and deployment. Below is a step-by-step breakdown:

1. Data Acquisition and Pre-processing:

- Data Reading and Initial Processing:

- Data was uploaded and read using Pandas in a Python environment, specifically from a CSV file named "churn.csv".
- Initial data overview included assessing the number of rows and columns, checking for missing values, and evaluating unique values for each feature.

- Data Cleaning and Transformation:

- The 'customerID' column, being a mere identifier, was dropped to ensure the model focused on relevant predictive features.
- Binary categorical features such as 'gender', 'SeniorCitizen', 'Partner', etc., were encoded into numerical format (0 and 1) using mapping functions.
- The 'SeniorCitizen' feature was specifically transformed from numerical (0, 1) to categorical ('No', 'Yes').
- Other categorical features with more than two categories were transformed using one-hot encoding via Pandas' `get_dummies` method.

- Feature Scaling:

- Features like 'tenure', 'MonthlyCharges', and 'TotalCharges' were scaled using MinMaxScaler to normalize their values.

2. Exploratory Data Analysis (EDA):

- Visual Exploration:

- Used Plotly for visual exploration of the target variable 'Churn' through pie charts.
- Developed a bar chart function for categorical variable analysis, visualizing the churn rate against different categorical features.
- Histograms were created for continuous variables to understand their distribution and relationship with churn.

3. Feature Engineering:

- Data Transformation and Feature Binning:

- Continuous variables were binned into categories (low, medium, high) for better trend analysis.
- A custom function was defined for histogram plotting of these binned variables.

4. Model Development:

- Baseline Model Creation:

- Various machine learning models like Logistic Regression, SVC, Random Forest, Decision Tree, and Naive Bayes were implemented.
- Performance metrics such as accuracy, precision, recall, and F1 score were used for model evaluation.
- The train-test split was performed with 70% training and 30% testing data.

- Feature Selection:

- Employed Recursive Feature Elimination with Cross-Validation (RFECV) to select the most relevant features for the Logistic Regression model.

5. Hyperparameter Tuning:

- Logistic Regression Optimization:

- Utilized RandomizedSearchCV for hyperparameter tuning of the Logistic Regression model.
- Defined a search space including parameters like 'solver', 'penalty', and regularization strength 'C'.

6. Final Model Evaluation and Deployment:

- Improved Model Performance Check:

- Retrained the Logistic Regression model using the best-found hyperparameters and evaluated its performance.
- The final model was saved using joblib for later use.

7. User Interface for Predictions:

- Interactive Prediction Interface:

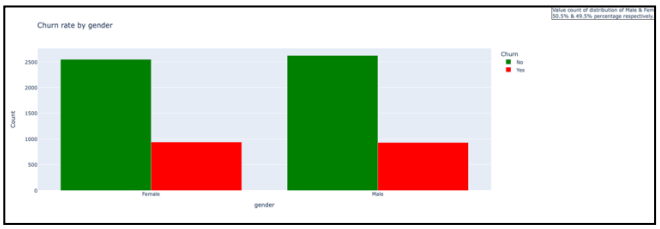
- Developed an interface using ipywidgets for user input.
- Created a preprocessing function to transform new input data for the model.
- Defined a process to collect user inputs, preprocess them, and use the saved model to predict churn.

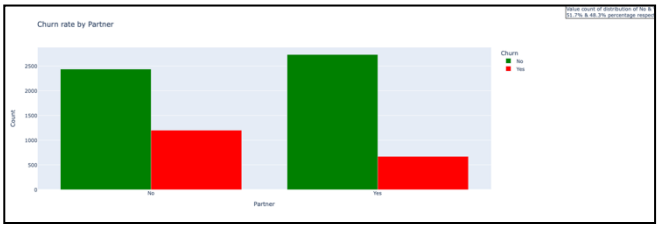
VI. Discussion and Analysis of Results

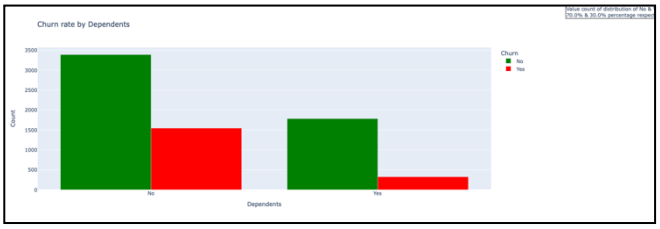
The project is focused on predicting customer churn for a telecom company using a dataset that includes customer information, usage patterns, and churn status. Within the dataset, 73.5% of customers discontinued the product, while the remaining 26.5% continued using it.

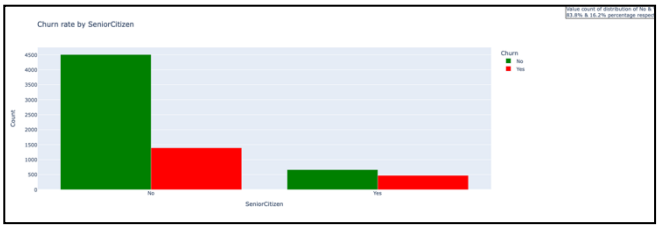
Demographic Analysis Insight: When examining churn rates based on gender, age, partner status, and department, it is observed that gender and partner distribution is even, with slight differences. Although there is a slightly higher churn

rate among females, the disparity is negligible. Younger customers (SeniorCitizen = No), those without partners, and those without dependents show a higher proportion of churn. The demographic analysis highlights non-senior citizens without partners and dependents as a specific segment with a higher likelihood of churning.

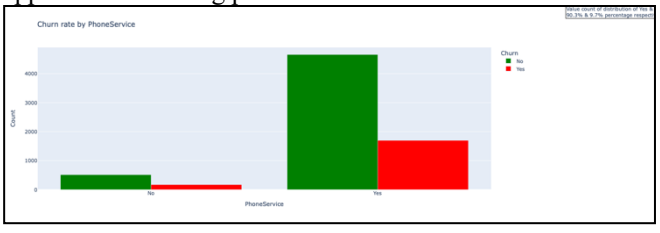


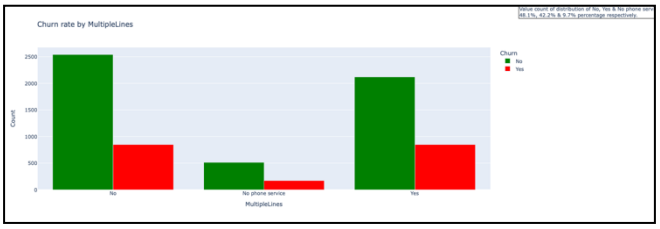


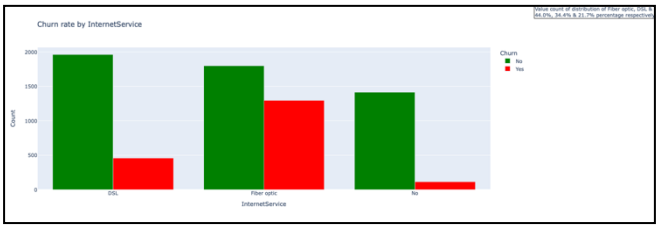


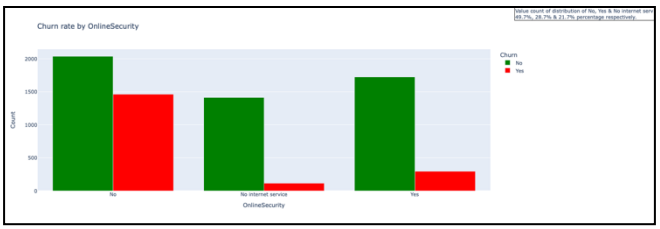


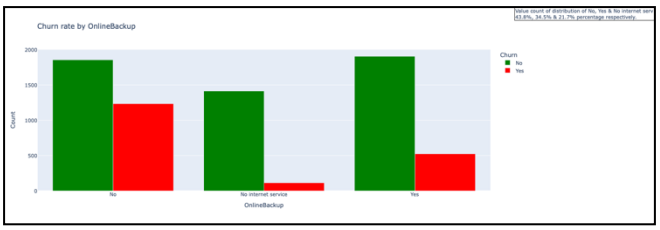
Services Subscribed by Customers Insight: Features related to services subscribed by customers exhibit significant variations. Customers without phone service are unable to have multiple lines. Approximately 90.3% of customers with phone services have a higher churn rate. Customers with fiber optic internet service are more likely to churn, possibly due to factors such as high prices or competition. Additional services like OnlineSecurity, OnlineBackup, DeviceProtection, and TechSupport are associated with lower churn rates. Streaming service does not appear to be a strong predictor of churn.

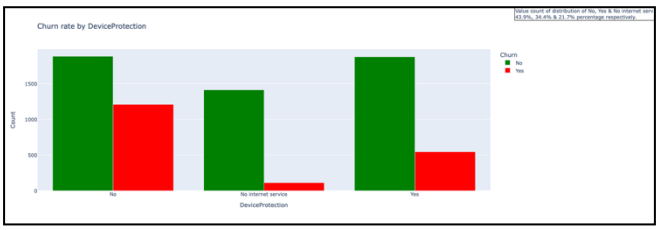


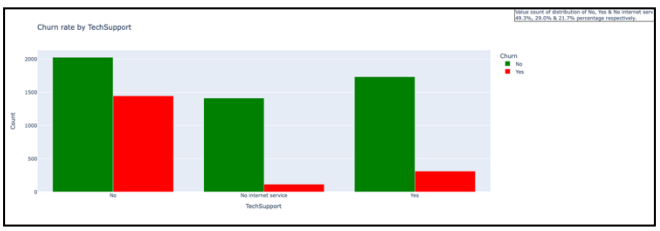


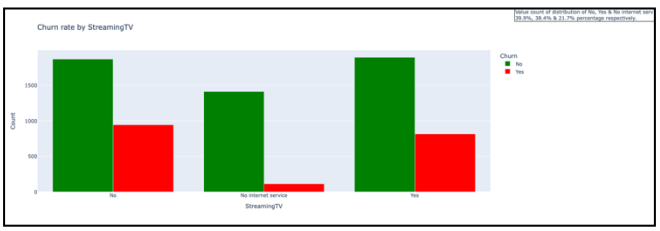


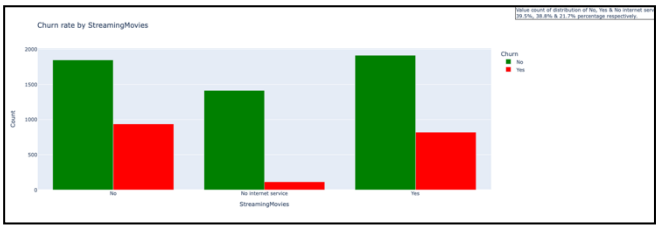




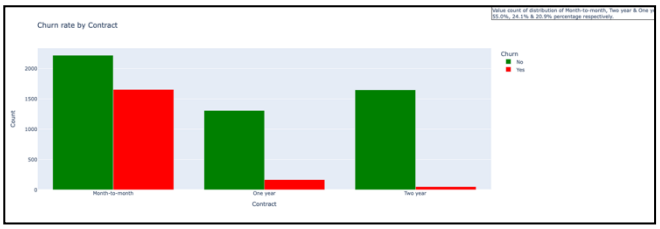


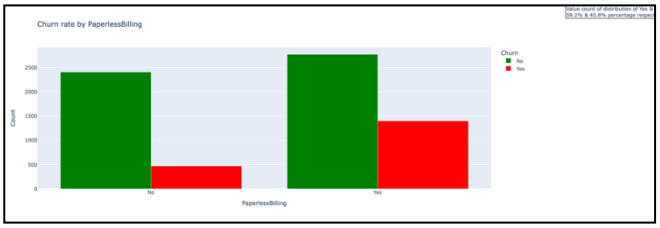


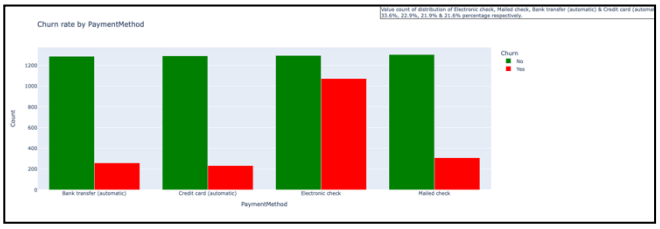




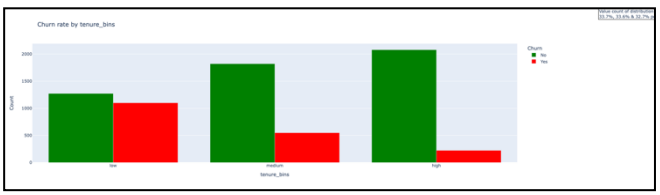
Payment Insights: A clear trend emerges where shorter contract durations correspond to higher churn rates, indicating that customers with longer plans face additional barriers to early cancellation. Higher churn rates are observed for customers opting for paperless billing, with 59.2% of customers using this billing method. Customers paying with electronic checks are more prone to churn compared to other payment types.

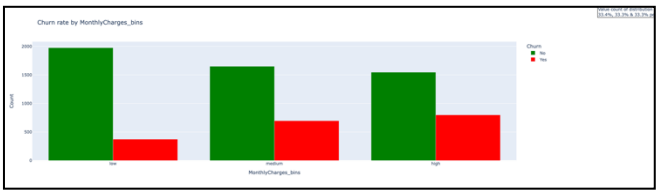


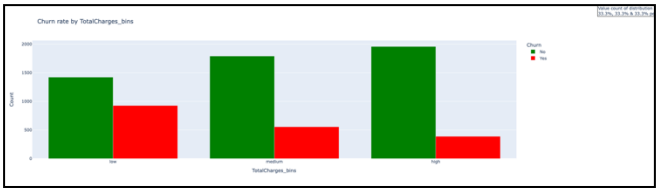




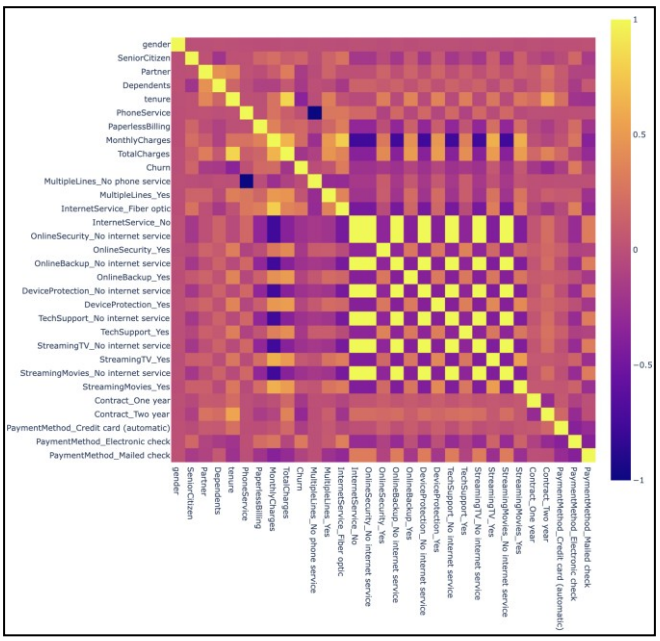
Customer Account Information Insight: The tenure histogram is right-skewed, indicating that most customers have been with the telecom company for the initial few months (0-9 months). The highest churn rate is also observed during this period. Notably, 75% of customers who churn do so within their first 30 months. The monthly charge histogram reveals that customers with higher monthly charges exhibit a higher churn rate, suggesting that discounts and promotions may influence customer retention.







Correlation: When dealing with high correlation between features, it is advisable to eliminate redundancy by dropping one of them. In this particular scenario, features like MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies, which exhibit high correlation, can be omitted.



The task of churn prediction is framed as a binary classification problem, where customers are categorized into either churned or retained within a specified timeframe. To guide the construction of an effective model, attention is directed towards addressing two key inquiries:

Identification of influential features for customer churn or retention: The significance of features is determined by examining the (P>|z|) column. A feature is deemed statistically significant if its absolute p-value is less than 0.05. Examples of such features include SeniorCitizen, Tenure, Contract, and PaperlessBillings.

Determination of the most impactful features for constructing a high-performance model: The second aspect revolves around assessing the importance of features, which is achieved by examining the exponential coefficient

values. These coefficients provide insights into the expected change in churn based on a one-unit alteration in a given feature.

VII. Conclusion

This methodology provided a robust framework for customer churn prediction. It involved meticulous data preprocessing, exploratory data analysis, feature engineering, model development and tuning, and finally, deployment with an interactive user interface for real-time predictions. This approach ensures that the model is not only accurate but also practical for real-world applications.

VIII. Future Works

Building on the current achievements, several areas of future work can be pursued to enhance the capabilities and applicability of the churn prediction model:

1. Advanced Machine Learning Techniques:

- Deep Learning Models: Investigate and implement deep learning architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) for more sophisticated feature extraction and pattern recognition.
- Ensemble Methods: Experiment with ensemble learning techniques like Gradient Boosting or Stacking to potentially improve prediction accuracy and model robustness.

2. Real-Time Prediction Capabilities:

- Develop and integrate real-time analytics to allow the model to respond to customer behavior and feedback dynamically, enhancing its applicability in fast-paced business environments.

3. Cross-Industry Validation and Adaptation:

- Test and adapt the model for use in industries beyond telecommunications, such as finance, retail, or e-commerce, to explore its generalizability and effectiveness in different business contexts.

4. Feature Engineering and Selection:

- Continuously update and refine the feature set to include new data sources and evolving customer behaviors.
- Implement automated feature engineering techniques to reduce manual effort and discover potentially predictive features.

5. Model Interpretability and Transparency:

- Focus on enhancing the interpretability of machine learning models, particularly if complex models are used, to ensure that stakeholders can understand and trust the model's predictions.

6. Handling Imbalanced Data:

- Explore advanced techniques for dealing with imbalanced datasets, such as different oversampling methods or advanced loss functions, to improve model performance, especially in predicting minority classes.

7. Integrating Customer Feedback Loop:

- Implement a system where customer feedback directly influences the model, allowing for continuous learning and adaptation to changing customer preferences and behaviors.

8. Personalization of Churn Predictions:

- Develop methods to personalize churn predictions and retention strategies for individual customers, enhancing the effectiveness of customer retention efforts.

9. Extensive Validation and Testing:

- Conduct extensive testing and validation of the model using various datasets to ensure reliability, including stress-testing the model under different scenarios.

10. Ethical Considerations and Bias Mitigation:

- Address potential ethical considerations, ensuring that the model does not inadvertently introduce or perpetuate biases, and is fair and equitable in its predictions.

By exploring these areas, the project can significantly advance the state-of-the-art in customer churn prediction, offering more robust, accurate, and user-friendly solutions for businesses aiming to enhance their customer retention strategies.

Acknowledgment

We express our profound gratitude to our esteemed guide, Professor Ashalatha Nayak, whose expertise and insights have been invaluable to the progression and success of this project. His unwavering support and constructive criticism have been pivotal in steering this research in the right direction.

We are also immensely thankful to the Department of Computer Science and Engineering at Manipal Institute of Technology for providing the necessary facilities and an enriching environment that fostered our intellectual growth and facilitated our research endeavours.

Our appreciation extends to our peers and the technical staff whose assistance and contributions have been instrumental throughout this journey. Their willingness to give their time so generously has been very much appreciated.

References

[1] Ahmed, A. A. E., Eltawil, A. B., & Hassanien, A. E. (2019). Using data mining techniques for customer churn prediction in telecom industry. International Journal of Advanced Computer Science and Applications, 10(6).

[2] Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276-286.

[3] Amin, A., Anwar, S., Adnan, A., Nawaz, M., et al. (2019). Machine learning for customer churn prediction: Algorithmic approaches and issues. ACM Computing Surveys (CSUR), 52(5), 1-36.

[4] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[5] Mozer, M. C., Wolniewicz, R., et al. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. IEEE Transactions on Neural Networks, 11(3), 690-696.

[6] K. Coussement and D. Van den Poel. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313-327.

[7] Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449.

[8] Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20), 7046-7056.

[9] Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515-524.

[10] Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, 34(10), 2902-2917