Data Mining Term Final

COMMERICAL INSURANCE CHURN

Melissa Young | DSC550 | 06/2023

INTRODUCTION

Customer retention plays a pivotal role in the long-term success of commercial insurance businesses. Understanding the factors that contribute to customer churn and implementing effective retention strategies are key priorities for insurers. This project aims to address the challenge of customer retention in the commercial insurance sector by leveraging a comprehensive dataset that encompasses vital information on business policies. By developing a predictive model, the objective is to empower insurance companies to proactively identify and address potential cancellations, ultimately enhancing customer retention rates and driving profitability.

The commercial insurance industry faces a significant problem in the form of a high rate of customer churn. Policy cancellations not only result in revenue loss but also hinder the establishment of lasting client relationships. To mitigate this issue, we aspire to develop a robust machine learning model capable of accurately predicting the likelihood of policy cancellation based on the various features available in the dataset.

The dataset comprises critical information such as policy label, effective date, expiration date, cancellation date, product code, subline code, street address, property city, property county name, property zip code, property state, written premium amount, total loss payments, and total reserve. By harnessing the power of advanced analytics, we can derive valuable insights from this data and construct a predictive model that helps insurance companies identify policyholders at risk of cancellation.

Through this predictive model, insurers can take proactive measures to address customer concerns, tailor retention strategies, and allocate resources effectively. By accurately assessing cancellation risks, insurance companies can engage in targeted communication, provide personalized solutions, and ultimately foster stronger relationships with their policyholders.

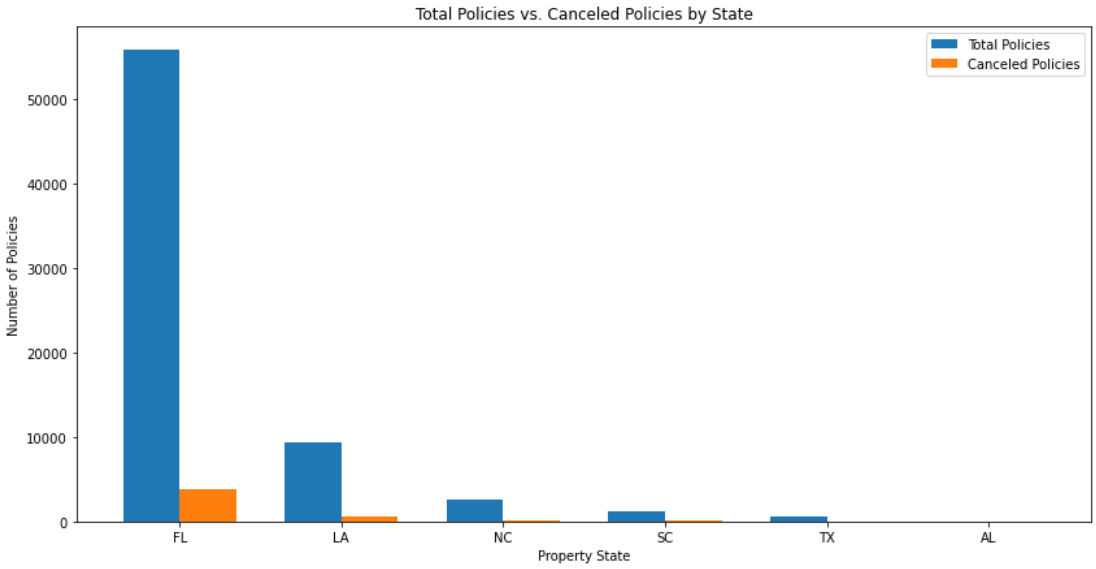
SUMMARY OF MILESTONES 1-3

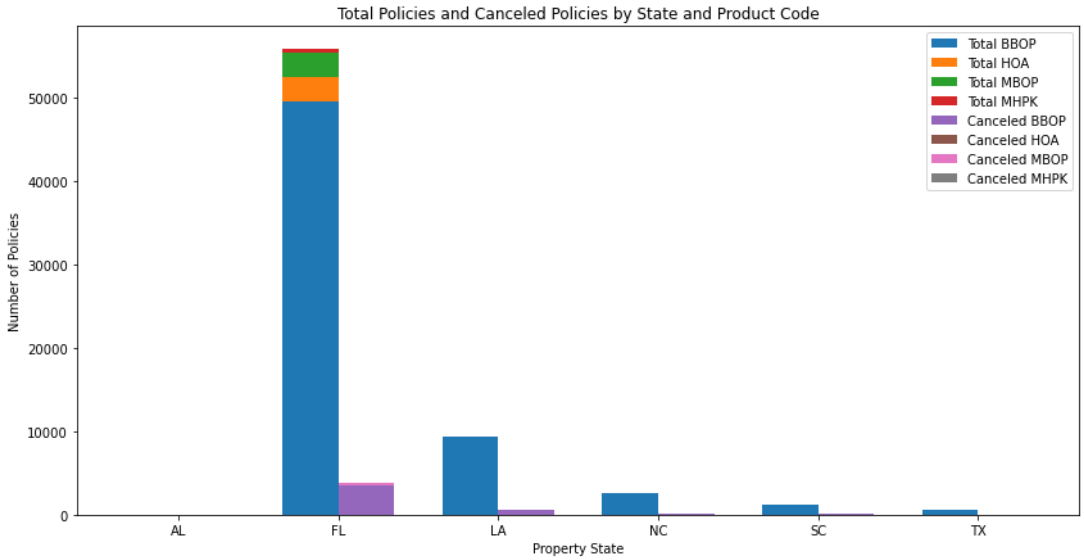
The target variable for our model is the "Cancellation Indicator," a binary variable derived from the cancellation date column. If a policy has a cancellation date, the indicator will be 1, signifying a cancellation. Otherwise, it will be 0, indicating that the policy is still active or expired naturally.

I started by cleaning and preprocessing the dataset, handling missing values, and removing duplicate records. Categorical variables will be converted into numerical variables using appropriate encoding techniques. I created new features, such as policy duration, time to cancellation, claim frequency, and loss ratio (total loss payments divided by written premium amount), which can provide additional insights into customer retention.

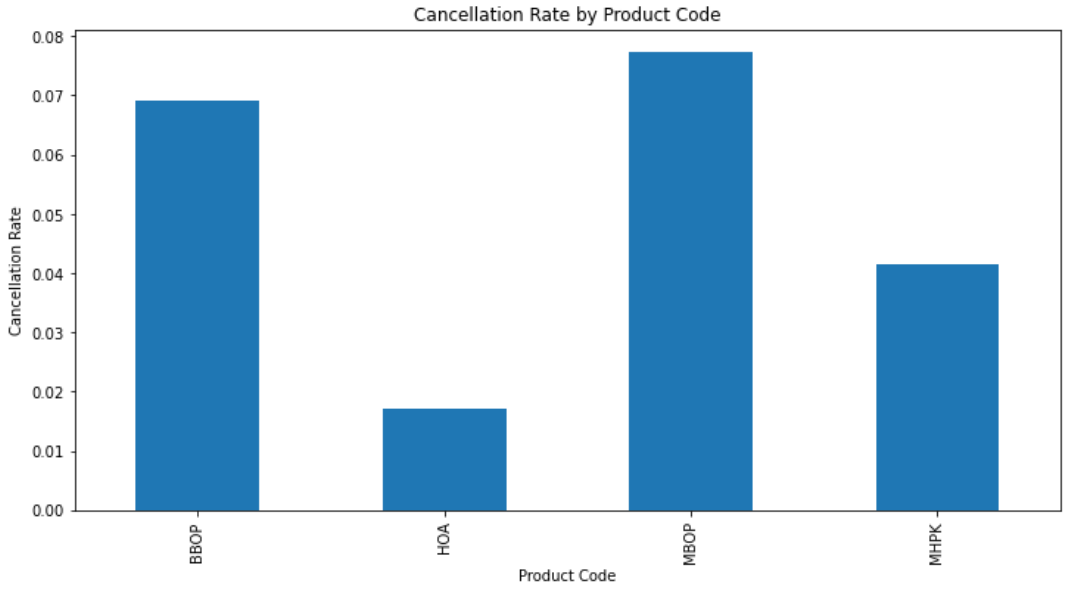
An extensive exploratory data analysis was performed to understand the relationships between different variables and cancellations. I identified patterns and trends in the data, such as geographical clusters of cancellations, seasonal variations, and correlations between specific product lines and cancellations.

The data utilized in this study was sourced from an internal database containing policy information. To ensure an adequate representation of the insurance landscape, a lookback period of three years was selected, resulting in a dataset encompassing 69,567 rows, corresponding to the number of policies sold during this timeframe. Upon identifying cancellations within the dataset, the remaining data comprised 4,656 rows. This indicates that approximately 7% of all policies written experienced cancellations prior to their expiration dates. It is noteworthy that the industry average for policy cancellations stands at less than 5%, underscoring the significance of this issue.

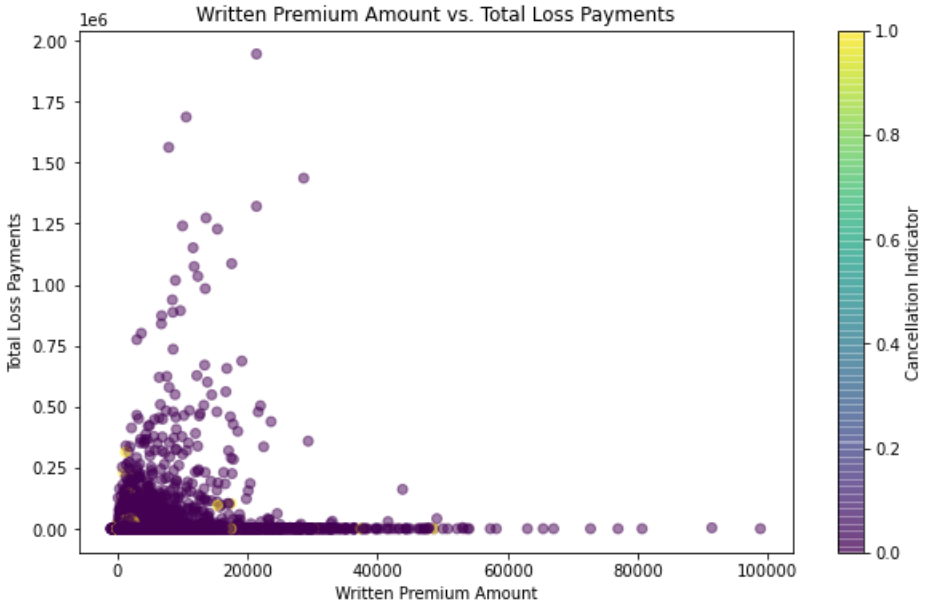




We can see that most of the policies written and canceled were in the states of Florida and Louisiana.



This bar chart shows the cancellation rate by product code further expanded. This visualization revealed patterns and trends related to specific product lines, which could clarify the factors that influence cancellations and improve customer retention strategies. As we can see here, MBOP and BBOP products have the highest potential for investigation.



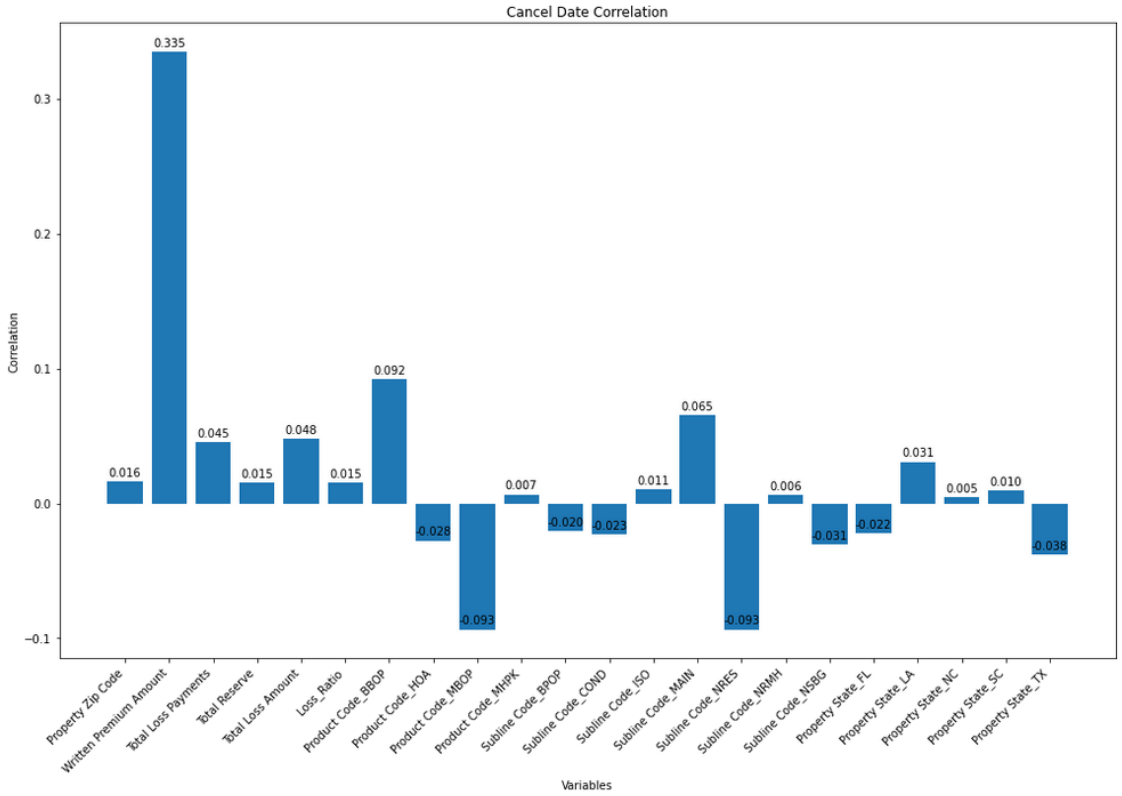
Written Premium Total: $180,240,321

Written Premium Canceled: $4,987,624

The above scatter plot shows the relationship between 'Written Premium Amount' and 'Total Loss Payments'. This scatter plot helped to visualize the correlation between the premium amounts and the loss payments for each policy.

The analysis revealed a loss indication of $5 million in premium due to these cancellations. This substantial financial impact emphasizes the urgency for insurance companies to tackle customer churn and implement effective retention strategies. By identifying the factors that contribute to policy cancellations, insurers can proactively address concerns, optimize their operations, and mitigate potential revenue losses.

Upon completing the EDA and cleaning of the datasets, it was decided to focus on cancellations only and dropped all columns that were not necessary to the model building. A correlation matrix was created to identify relationships pertaining to the number of days between policy start date and cancellation date, an indicator which I chose to call “Cancel Date Difference Days”. Below is a graph showing output with the highest correlation with Written Premium Amount.



To optimize the modeling process and enhance the accuracy of predictions, a feature selection step was performed based on correlation analysis. The features with the least correlation to the target variable were dropped from the dataset, ensuring that only the most relevant and influential features remained for the modeling stage.

Following the feature selection step, the dataset was divided into training and test sets using the widely adopted 80/20 rule. This partitioning ensured that 80% of the data was allocated for training the model, allowing it to learn patterns and relationships within the data. The remaining 20% of the data was reserved for evaluating the model's performance and assessing its ability to generalize to unseen data.

I first ran a linear regression model that did not perform well with the results:

```
Training data R2, RMSE, MAE:
R2: 0.1281354607918873
RMSE: 100.80555402420825
MAE: 84.30556881731694

Test data R2, RMSE, MAE:
R2: 0.04869079546943145
RMSE: 104.22555587997634
MAE: 84.33839530373024
```

Based on these metrics, it appears that the linear regression model may not be capturing the underlying patterns in the data very well. The R2 values are relatively low, indicating that the model explains only a small portion of the variance in the target variable. Additionally, the RMSE and MAE values are relatively high, suggesting that the model's predictions have a significant average error.

I began to experiment with other models to get best fit.

Decision Tree output without hyperparameters:

```
Training data R2, RMSE, MAE:
R2: 0.9997617960517985
RMSE: 1.666228012450182
MAE: 0.06015037593984962

Test data R2, RMSE, MAE:
R2: -0.05207293066791996
RMSE: 109.60649597185136
MAE: 78.2043991416309
```

The high R2 value (close to 1) and low RMSE and MAE values for the training data indicate that the decision tree model fits the training data very well. However, the negative R2 value and relatively high RMSE and MAE values for the test data suggest that the model does not generalize well to unseen data. This indicates potential overfitting, where the model has learned the training data too well and does not perform well on new data.

Decision Tree output with hyperparameters:

```
Best Hyperparameters: {'max_depth': 5, 'min_samples_leaf': 4,
'min_samples_split': 2}

Best Model Performance:
R2: 0.46985001203827836
RMSE: 77.80593904371312
MAE: 58.124266892072825
```

These metrics indicate that the model with tuned hyperparameters performs better than the previous model. The R2 score of 0.47 suggests that the model explains 47% of the variance in the target variable. The lower values of RMSE and MAE indicate smaller prediction errors.

I then ran a random forest model without hyperparameters:

```
Training data R2, RMSE, MAE:
R2: 0.9192363650297426
RMSE: 30.680884965184404
MAE: 22.669396053910287
```

```
Test data R2, RMSE, MAE:
R2: 0.41774800894869635
RMSE: 81.53965354055921
MAE: 59.885281677907216
```

Overall, the Random Forest model shows good performance on the training data, but there is some drop in performance on the test data. This may indicate a degree of overfitting, where the model is capturing the training data patterns too closely and not generalizing well to new data. Further optimization or tuning of the model parameters could potentially improve its performance on the test data.

Random forest model with hyperparameters:

```
Best Hyperparameters: {'max_depth': 5, 'min_samples_leaf': 4,
'min_samples_split': 2, 'n_estimators': 300}

Best Model Performance (Training data):
R2: 0.5131996887112908
RMSE: 75.32434734086904
MAE: 57.03227118231391

Best Model Performance (Test data):
R2: 0.47818243711817887
RMSE: 77.19207524240134
MAE: 57.37015112400198
```

The results indicate improved performance compared to the previous iterations.

CONCLUSION

In conclusion, the results of the analysis demonstrate that the best Random Forest model, with tuned hyperparameters, exhibited improved performance in predicting the "Cancel Date Difference Days" based on the given features. It achieved an R2 score of 0.48 on the test data, indicating that it explains 48% of the variance in the target variable. Additionally, the relatively small prediction errors, as evidenced by the RMSE value of 77.19 and MAE value of 57.37, further emphasize the model's effectiveness.

The performance of this model compared to previous iterations showcases its potential as a reliable predictor of cancellation patterns. However, there is still room for improvement, and incorporating additional features could enhance the predictive capabilities further.

One potential area for improvement is to incorporate more customer-related demographics, such as age, gender, location, or income level. By including these factors, we may uncover hidden patterns and associations that contribute to cancellation behaviors.

Furthermore, considering customer behaviors, such as payment history, can provide valuable insights into cancellation patterns. Examining variables like payment frequency, punctuality, or payment method could reveal correlations that contribute to the likelihood of cancellations.

In addition to customer demographics and behaviors, analyzing cancellation reasons through textual analysis can provide a deeper understanding of the underlying factors that influence cancellations. By exploring the language and sentiment used in cancellation requests or feedback, we can gain insights into specific pain points or issues that drive customers to cancel.

Incorporating these new features and conducting further analysis has the potential to significantly improve the accuracy and effectiveness of the cancellation prediction model. By continuously refining and expanding the model's capabilities, we can make more informed decisions and take proactive measures to mitigate cancellations in the future.