ALY 6020 – CRN 70409

Predictive Analytics

Professor: Amin Karimpour Northeastern University, Boston



Module 6 Assignment – Final Exam

Submitted by – Santhi Kiran Chavali

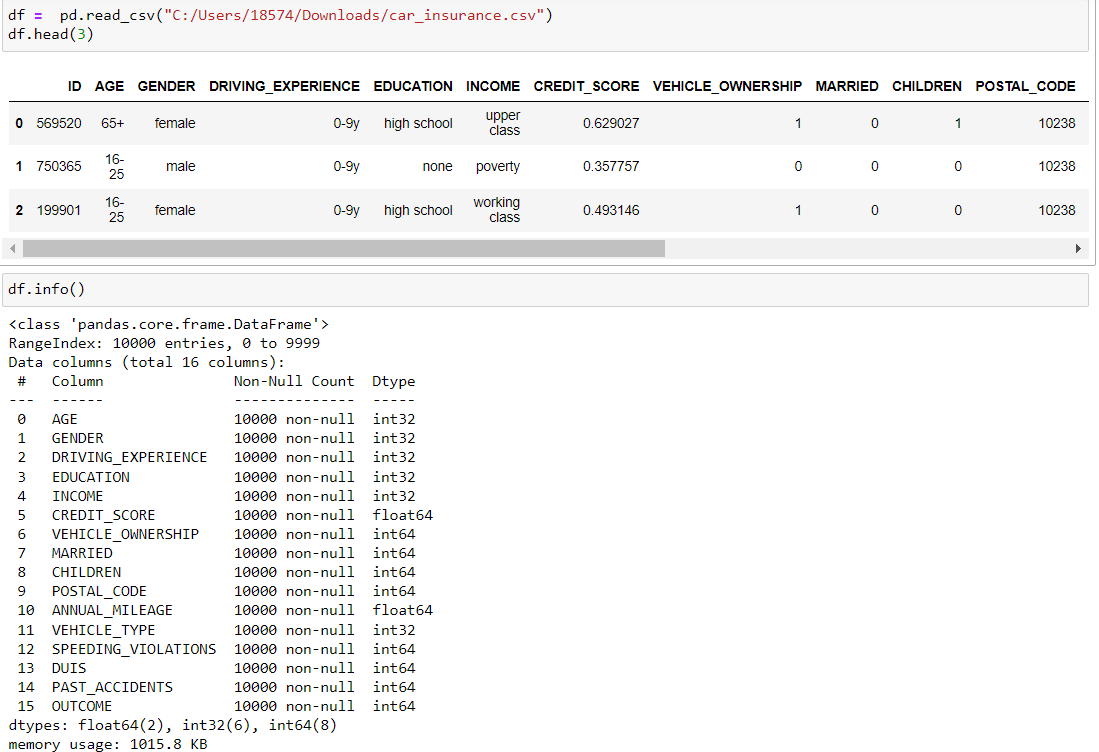
**Predictive Modeling for Car Insurance Risk Assessment and Policy Pricing Optimization**

**Introduction:**

For this project, we are presented with a car insurance dataset that contains information about policyholders and their vehicles, specifically whether or not they were involved in a car accident within a given year. The goal of this assignment is to explore and analyze this dataset to better understand the factors that contribute to car accidents. Our primary objective is to build predictive models that can identify significant variables influencing accident risk and to provide recommendations for an insurance company based on our findings. The assignment involves several key steps which include cleaning the dataset, addressing missing values, and preparing the data for analysis. We employ non-ensemble and ensemble models, such as Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and a Neural Network, to predict car accidents. We evaluate their accuracy and identify significant variables. We construct confusion matrices to analyze the performance of each model, specifically focusing on reducing False Positives. Based on the model results, we provide specific recommendations to the insurance company on what to consider when applicants apply for car insurance and how to reduce the risk associated with policyholders.

# Analysis:

In the initial steps of the analysis, we imported several essential libraries into Jupyter Notebook to facilitate data exploration and modeling. These libraries included Pandas for data manipulation, NumPy for numerical operations, Seaborn for data visualization, and Scikit-Learn for machine learning tasks. We also mentioned the inclusion of the statistics library, which is likely for future statistical analysis. The next crucial step involved importing the dataset into Jupyter Notebook. Upon inspecting the dataset, we observed that it contains a total of 10,000 records with 16 columns, providing a substantial amount of data for analysis. This dataset would serve as the basis for our investigation into car insurance and accident prediction.

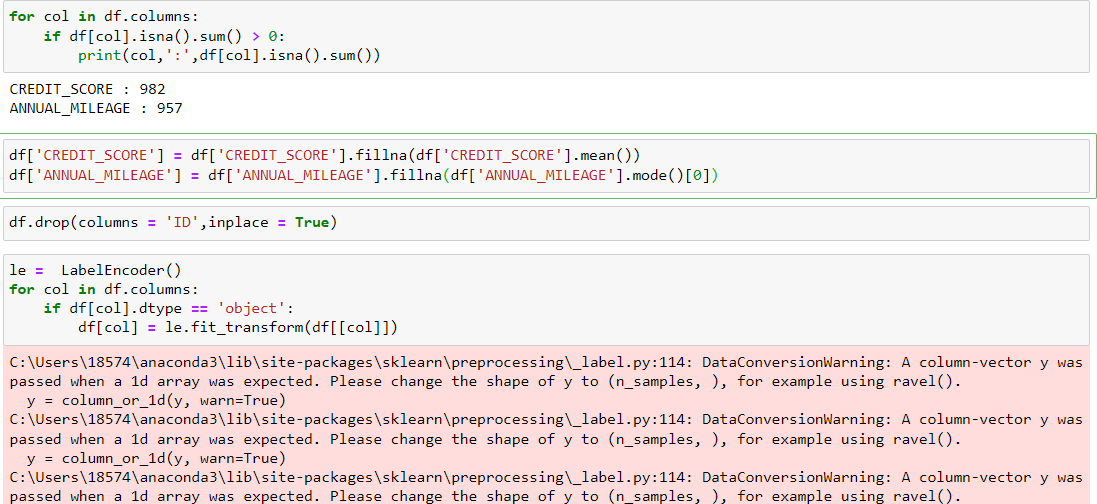


One of the first data preprocessing tasks was to identify and address missing values. During this process, we discovered that the 'CREDIT\_SCORE' column had 982 missing values, while the 'ANNUAL\_MILEAGE' column had 957 missing values. To ensure the integrity of our analysis, we took the following actions:

* We replaced the missing values in the 'CREDIT\_SCORE' column with the mean value of the column. This imputation strategy helped maintain the statistical characteristics of the data and ensured that missing values did not adversely impact our analysis.
* For the 'ANNUAL\_MILEAGE' column, we replaced missing values with the mode, the most frequently occurring value. This decision was made based on the assumption that the mode would be a representative estimate for missing annual mileage values. By addressing these missing values, we prepared the dataset for further analysis.

Additionally, the 'ID' column, which was found to have no practical use for subsequent analysis, was dropped. This column likely represented a unique identifier for each policyholder but did not

provide meaningful information for our modeling and predictions. Lastly, to enable the use of categorical variables in machine learning models, we employed the Label Encoder to convert categorical values into numerical representations. This transformation ensured that the models could work with all data types effectively and was a fundamental step in our data preparation process.

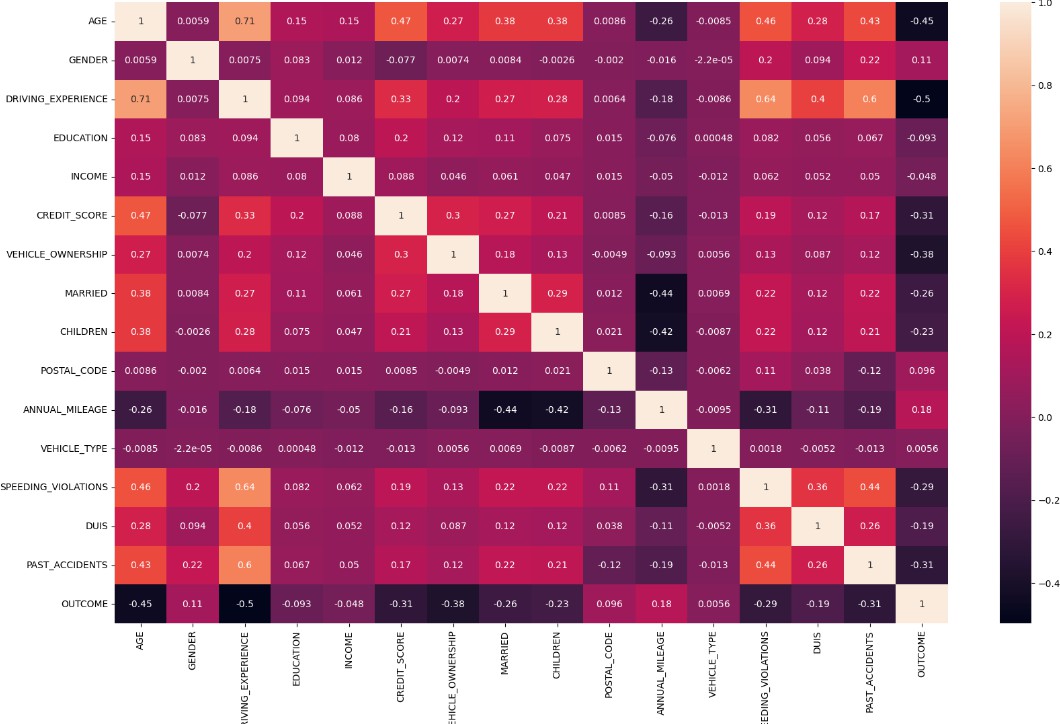


These initial steps laid the foundation for our comprehensive analysis of the car insurance dataset and subsequent modeling to predict accident risk. By addressing missing values, dropping unnecessary columns, and encoding categorical variables, we prepared the dataset for further exploration and the development of predictive models.

# Correlation:

In the course of our analysis, we conducted a correlation analysis to better understand the relationships between different variables within the dataset. Correlation measures the degree to which two variables are linearly related, providing insights into potential associations and dependencies between features. Here are some key observations from our correlation analysis:

Correlation Heatmap: We created a correlation heatmap using Seaborn to visualize the correlation coefficients between pairs of variables in the dataset. This heatmap allowed us to quickly identify strong positive and negative correlations.



Variable Relationships: We observed several interesting relationships between variables. For instance, 'AGE' and 'DRIVING\_EXPERIENCE' showed a strong positve correlation, which is intuitive since older individuals tend to have more driving experience. 'PAST\_ACCIDENTS' and 'SPEEDING\_VIOLATIONS' displayed a positive correlation, indicating that individuals with a history of accidents tend to have more speeding violations.

Target Variable: We also examined the correlations of each variable with the target variable, 'OUTCOME,' which represents whether a policyholder had an accident that year. This analysis revealed the influence of each feature on accident predictions. Notably, 'CREDIT\_SCORE' exhibited a weak negative correlation with 'OUTCOME,' suggesting that a higher credit score might be associated with a lower likelihood of having an accident.

# Logistic Regression:

In our analysis, we built a logistic regression model to predict car accidents, providing valuable insights into the significance of different variables. This model helps identify significant factors and their corresponding coefficients, enabling us to understand their impact on accident predictions. Logistic regression is essential for assessing the influence of each feature on the likelihood of an accident.

# Significant Variables and Their Coefficients:

|  |  |
| --- | --- |
| Variable | Coefficient |
| AGE | -0.31 |
| GENDER | 0.01 |
| DRIVING\_EXPERIENCE | -0.31 |
| EDUCATION | -0.08 |
| INCOME | -0.08 |
| CREDIT\_SCORE | -0.04 |
| VEHICLE\_OWNERSHIP | -0.12 |
| MARRIED | -0.10 |
| CHILDREN | -0.09 |
| POSTAL\_CODE | 0.00 |
| ANNUAL\_MILEAGE | 0.00 |
| VEHICLE\_TYPE | -0.00 |
| SPEEDING\_VIOLATIONS | -0.44 |
| DUIS | -0.07 |
| PAST\_ACCIDENTS | -0.33 |

The logistic regression model identified the variables above as significant, each with an associated coefficient. Notably, 'SPEEDING\_VIOLATIONS' and 'PAST\_ACCIDENTS' exhibited the most substantial negative coefficients, suggesting that a history of speeding violations and past accidents is strongly associated with an increased risk of car accidents. These findings align with common intuition and demonstrate the predictive power of these variables.

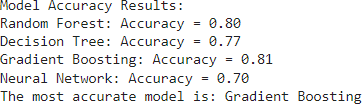
The model aids the insurer in risk assessment for tailored policies and premiums, adjusting pricing based on risk, and promoting safe driving through incentive programs. It also enhances claims management, reducing financial losses.

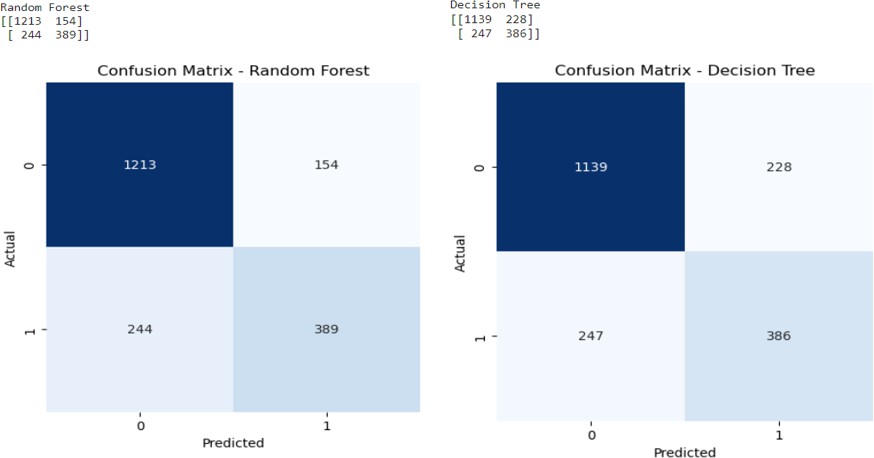
Model Accuracy:

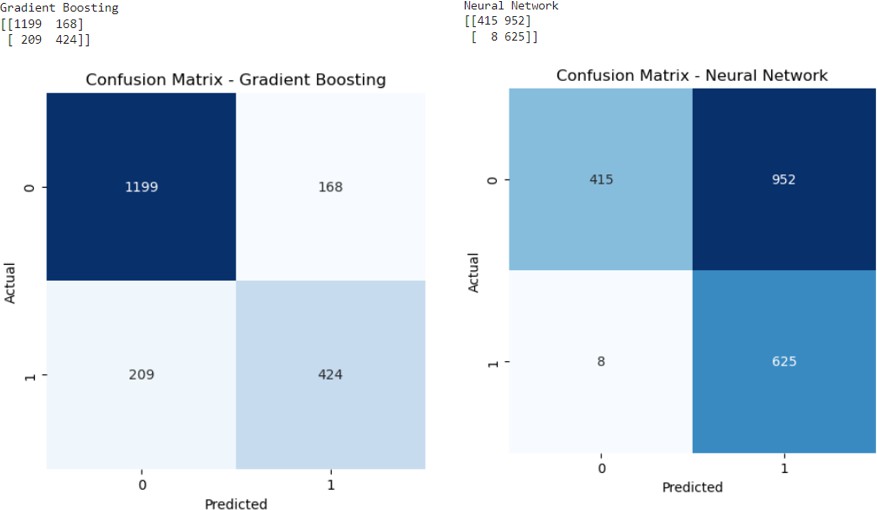
The logistic regression model achieved an accuracy of 0.79, indicating its effectiveness in predicting accidents. This accuracy, combined with the knowledge of significant variables, empowers the insurer to make informed decisions about policyholders' risk and pricing.

# Running Non-Ensemble Models and Analyzing Confusion Matrices:

In this phase of our analysis, we implemented non-ensemble models commonly used in class: Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, and a Neural Network. Each model was trained with 1000 iterations to predict car accidents. We assessed their accuracy and the reasoning behind choosing these models. We also built confusion matrices to understand their performance and discussed which part of the matrix an insurance company would want to reduce and which model is most effective in achieving this goal.







# Model Accuracy Results:

The accuracy results of the models are as follows: Logistic Regression: Accuracy = 0.79

Random Forest: Accuracy = 0.80 Decision Tree: Accuracy = 0.77

**Gradient Boosting: Accuracy = 0.81** Neural Network: Accuracy = 0.70 **Model Selection and Reasons:**

Among the models, Gradient Boosting demonstrated the highest accuracy, making it the most suitable choice for accurate accident prediction. Gradient Boosting excels in combining weak learners to create a strong predictive model and has proven its effectiveness in numerous real- world applications.

# Confusion Matrices and Reducing False Positives:

We constructed confusion matrices for each model to assess their performance. A crucial aspect for an insurance company is reducing False Positives (FP), as these errors result in incorrect accident predictions, potentially leading to higher insurance costs for policyholders.

Random Forest: TP = 389, TN = 1213, FP = 154, FN = 244. Decision Tree: TP = 386, TN = 1139, FP = 228, FN = 247.

Gradient Boosting: TP = 424, TN = 1199, FP = 168, FN = 209. Neural Network: TP = 625, TN = 415, FP = 952, FN = 8.

Among the models, Gradient Boosting exhibited the best performance in reducing False Positives, making it the most effective in minimizing incorrect accident predictions. This is particularly valuable for the insurance company, as it ensures more accurate risk assessments and potential cost savings for policyholders.

# Important Features and Recommendations:

From our analysis, several features emerged as vital for predicting car accidents. Notably, 'SPEEDING\_VIOLATIONS' and 'PAST\_ACCIDENTS' proved to be highly significant in all models, with

strong negative coefficients, signifying a direct correlation between a history of speeding violations and past accidents with an increased risk of accidents. These findings underscore the importance of monitoring and addressing risky driving behaviors.

Based on these insights, we recommend that the insurance company closely examines these factors when individuals apply for car insurance. It's crucial to consider an applicant's history of speeding violations and past accidents when assessing their risk profile. High-risk applicants could be subject to higher premiums or safety requirements, such as attending defensive driving courses.

Furthermore, the insurance company can play an active role in reducing risk for policyholders. Offering incentives for safe driving practices, such as discounts for maintaining a clean driving record, can be a powerful motivator. By promoting responsible behavior on the road and offering guidance on risk reduction, the company can contribute to safer roads, reduced accidents, and enhanced customer satisfaction. This approach not only minimizes claims but also fosters a culture of responsible driving, benefiting both policyholders and the insurer.

# Conclusion:

In this comprehensive analysis of a car insurance dataset, we successfully built predictive models and explored critical factors that influence the likelihood of car accidents. Our primary objective was to assist an insurance company in making data-driven decisions related to risk assessment, policy pricing, and accident prevention. Here are the key takeaways:

**Predictive Models:** We implemented various models, including logistic regression, random forest, decision tree, gradient boosting, and a neural network, to predict car accidents. Gradient Boosting emerged as the most accurate model with an accuracy of 0.81.

**Significant Variables:** We identified key variables that play a crucial role in predicting accidents. 'SPEEDING\_VIOLATIONS' and 'PAST\_ACCIDENTS' were consistently significant across models, indicating that a history of risky driving behaviors substantially increases accident risk.

**Insurance Recommendations:** We recommended that the insurance company closely scrutinizes an applicant's history of speeding violations and past accidents when assessing their risk profile.

Higher-risk applicants should face higher premiums or safety requirements, such as defensive driving courses. Furthermore, promoting safe driving practices through incentive programs can encourage policyholders to drive responsibly and maintain clean records.

**Risk Reduction:** By understanding the significance of these variables, the insurer can contribute to safer roads, lower accident rates, and reduced financial losses. The knowledge gained from our analysis empowers the company to make informed decisions and enhance its risk assessment and customer satisfaction.

In conclusion, our analysis provides actionable insights for the insurance company to better assess and manage risk, ultimately resulting in more precise policies, improved road safety, and increased customer loyalty. By leveraging these findings, the insurer can foster a safer and more responsible driving environment while optimizing its business operations

# References:

Choksi, Y. (2020, March 9). Model selection: Cp, AIC, BIC and adjusted R2. Analytics Vidhya. [https://medium.com/analytics-vidhya/model-selection-cp-aic-bic-and-adjusted-r2-](https://medium.com/analytics-vidhya/model-selection-cp-aic-bic-and-adjusted-r2-6a0af25945b6)

[6a0af25945b6](https://medium.com/analytics-vidhya/model-selection-cp-aic-bic-and-adjusted-r2-6a0af25945b6)

Gupta, D. (2020, January 29). Activation Functions | Fundamentals Of Deep Learning. Analytics Vidhya. [https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-](https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/) [activation-functions-when-to-use-them/](https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/)

Difference between random forest and Gradient boosting Algo. (n.d.). Www.linkedin.com. Retrieved October 21, 2023, from [https://www.linkedin.com/pulse/difference-between-random-](https://www.linkedin.com/pulse/difference-between-random-forest-gradient-boosting-algo-goon/) [forest-gradient-boosting-algo-goon/](https://www.linkedin.com/pulse/difference-between-random-forest-gradient-boosting-algo-goon/)

Pratt, M. (2020, June 9). 10 Common Uses for Machine Learning Applications in Business. SearchEnterpriseAI. [https://www.techtarget.com/searchenterpriseai/feature/10-common-uses-](https://www.techtarget.com/searchenterpriseai/feature/10-common-uses-for-machine-learning-applications-in-business) [for-machine-learning-applications-in-business](https://www.techtarget.com/searchenterpriseai/feature/10-common-uses-for-machine-learning-applications-in-business)

Sagar, R. (2020, July 15). Top Optimisation Methods In Machine Learning. Analytics India Magazine. [https://analyticsindiamag.com/optimisation-machine-learning-methods-gradient-](https://analyticsindiamag.com/optimisation-machine-learning-methods-gradient-descent/) [descent/](https://analyticsindiamag.com/optimisation-machine-learning-methods-gradient-descent/)

Sadananda, N. (2021, September 29). The battle between Logistic Regression, Random Forest Classifier, XG Boost and Support Vector…. Medium. [https://medium.com/@nischitasadananda/the-battle-between-logistic-regression-random-](https://medium.com/%40nischitasadananda/the-battle-between-logistic-regression-random-forest-classifier-xg-boost-and-support-vector-46d773c70f41) [forest-classifier-xg-boost-and-support-vector-46d773c70f41](https://medium.com/%40nischitasadananda/the-battle-between-logistic-regression-random-forest-classifier-xg-boost-and-support-vector-46d773c70f41)

Mahesh. (2023, June 13). Exploring Decision Trees, Random Forest, Logistic Regression, KNN, Linear Regression, SVM, RNN…. Medium. [https://medium.com/@maheshhkanagavell/exploring-](https://medium.com/%40maheshhkanagavell/exploring-) decision-trees-random-forest-logistic-regression-knn-linear-regression-svm-rnn-bf52fa94a066