Predictive Model for Identifying 'Bad Buys' at Auto Auctions

Introduction US used-car-dealership company buys used-cars cheaply in online auctions and from other car sellers in order to then resell them profitably on their own platform. It's not always easy to tell if a used car is worth buying: one of the biggest challenges with used car auctions is the risk of a car having serious problems that prevent it from being resold to customers. These are so-called "lemon cars" - that is, cars that come out of the box with significant defects due to manufacturing errors that significantly affect the safety, use, or value of that car, and at the same time not in a reasonable number of repairs or within a given period can be remedied. In this case, the customer has the right to have the purchase price refunded. In addition to the acquisition costs, the wrong purchase of such "lemon cars" leads to considerable follow-up costs, such as the storage and repair of the car, which can result in losses when the vehicle is resold. That's why it's important to the company to rule out as many bad purchases of "lemon cars" as possible. In order to relieve the buyers in the company with the huge number of offers, a model should be developed that predicts whether an offer would be a bad buy in the sense of a lemon car. However, this must not lead to too many good purchases being excluded.

Industry Review

Market Size and Growth

- The used car market is substantial, with significant growth due to increasing consumer demand for affordable vehicles.
- In the U.S., the market size is over $800 billion, with millions of used cars sold annually.

Consumer Trends

- Consumers prefer used cars due to economic factors, lower prices, and certified pre-owned programs.
- Online platforms and digital marketplaces provide more options and transparency for consumers.

Auction Landscape

- Auto auctions are crucial for dealerships to acquire inventory.
- Major auction companies like Manheim and ADESA dominate, offering both physical and online auction services.

Technological Advancements

- Digital platforms enable online bidding and virtual inspections.
- Machine learning and predictive analytics assess vehicle conditions and predict resale values.

Vehicle Inspection and Grading

- Comprehensive inspections and standardized grading systems evaluate the condition of used cars.
- Inspections assess mechanical condition, appearance, and performance, providing transparency to buyers.

Economic Factors

- The used car market is influenced by interest rates, unemployment rates, and disposable income levels.
- Economic downturns increase demand for used cars as cost-effective transportation options.

Regulatory Environment

- The industry is regulated by emissions standards, safety requirements, and consumer protection laws.
- Compliance is crucial for avoiding legal issues and maintaining consumer trust.

Supply Chain Dynamics

- The supply of used cars is affected by lease returns, trade-ins, and fleet liquidations.

Data Collection

Content:

- Sources of data (e.g., auction transactions, vehicle inspections)
- Types of data collected (e.g., vehicle age, mileage, condition)
- Data collection methods and tools

Overview of the entire project lifecycle from data collection to deriving insights

Data Cleaning

Content:

- Handling missing values
- Removing duplicates
- Correcting inaccuracies
- Standardizing formats

Exploratory Data Analysis

Content:

- Key findings from initial data analysis
- Visualizations (e.g., histograms, scatter plots)
- Identified trends and patterns

Data Preprocessing

- The goal of data preparation is to find a way to clean the data sets for our model (data cleaning) and to bring them into a format that the model can read (data type transformation).
- When these steps have been completed, we can set about selecting a training set that is as representative as possible (sampling).

- Feature engineering (creating new variables)
- Scaling and normalization
- Handling categorical data (e.g., encoding)

Model Building

- Description of chosen models (e.g., logistic regression, random forest)
- Rationale for model selection
- Implementation details

Model Evaluation Techniques

Content:

- Metrics used for evaluation (e.g., accuracy, precision, recall)
- Cross-validation
- Results of model evaluation

Deriving Business Metrics and Business Insights

Content:

- Key metrics (e.g., cost savings, reduction in 'bad buys')
- Insights gained from the model
- Implications for dealership operations and decision-making

Conclusion and Future Steps

Content:

- Summary of key findings and results
- Next steps for further improvement (e.g., model refinement, additional data sources)
- Potential impact on the industry

Modelling Approach

- Data splitting: 80% train, 20% test
- Models to evaluate: Logistic Regression (baseline)
- Random Forest
- KNeighborsClassifier
- GridSearchCV
- Handling class imbalance:
- Hyperparameter tuning
- Class weighting
- Evaluation metrics:
- AUC-ROC, precision, recall, F1-score

Improvements

1. **Alternative Data Imputation**: Instead of using the median for numerical value imputation, try other methods such as k-Nearest Neighbors (kNN) or the most frequent value to handle missing data more effectively.
2. **Advanced Resampling Techniques**: Apply techniques like Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, as duplicating minority class records often fails to add new information.
3. **Ensembling Models**: Utilize ensemble methods like VotingClassifier from scikit-learn to combine different models. This approach aims to balance high variance (quick adaptation to changes) and low bias (stability) for improved predictions.

4. **Hyperparameter Optimization**: Explore different combinations and optimal sets of hyperparameters beyond grid search to enhance model performance.

Conclusion:

In conclusion, this assignment aimed to develop a predictive model to identify "lemon cars" from auction and seller listings to mitigate the risks and costs associated with purchasing defective vehicles. Using recall, precision, and F1 score metrics, focused on maximizing recall to exclude as many lemon cars as possible while maintaining a low false positive rate.

Through comprehensive exploratory data analysis (EDA), identified key factors influencing the likelihood of a vehicle being a lemon, such as vehicle age, manufacturer, and acquisition cost.

This model effectively distinguished between good and bad purchases, demonstrating strong predictive capabilities. The findings underscore the importance of data-driven approaches in enhancing decision-making processes within the used-car industry. Implementing this model can significantly reduce the incidence of bad buys, leading to improved profitability and customer satisfaction for the dealership company. Future work could explore integrating additional data sources and advanced machine learning techniques to further refine the model's accuracy and robustness.