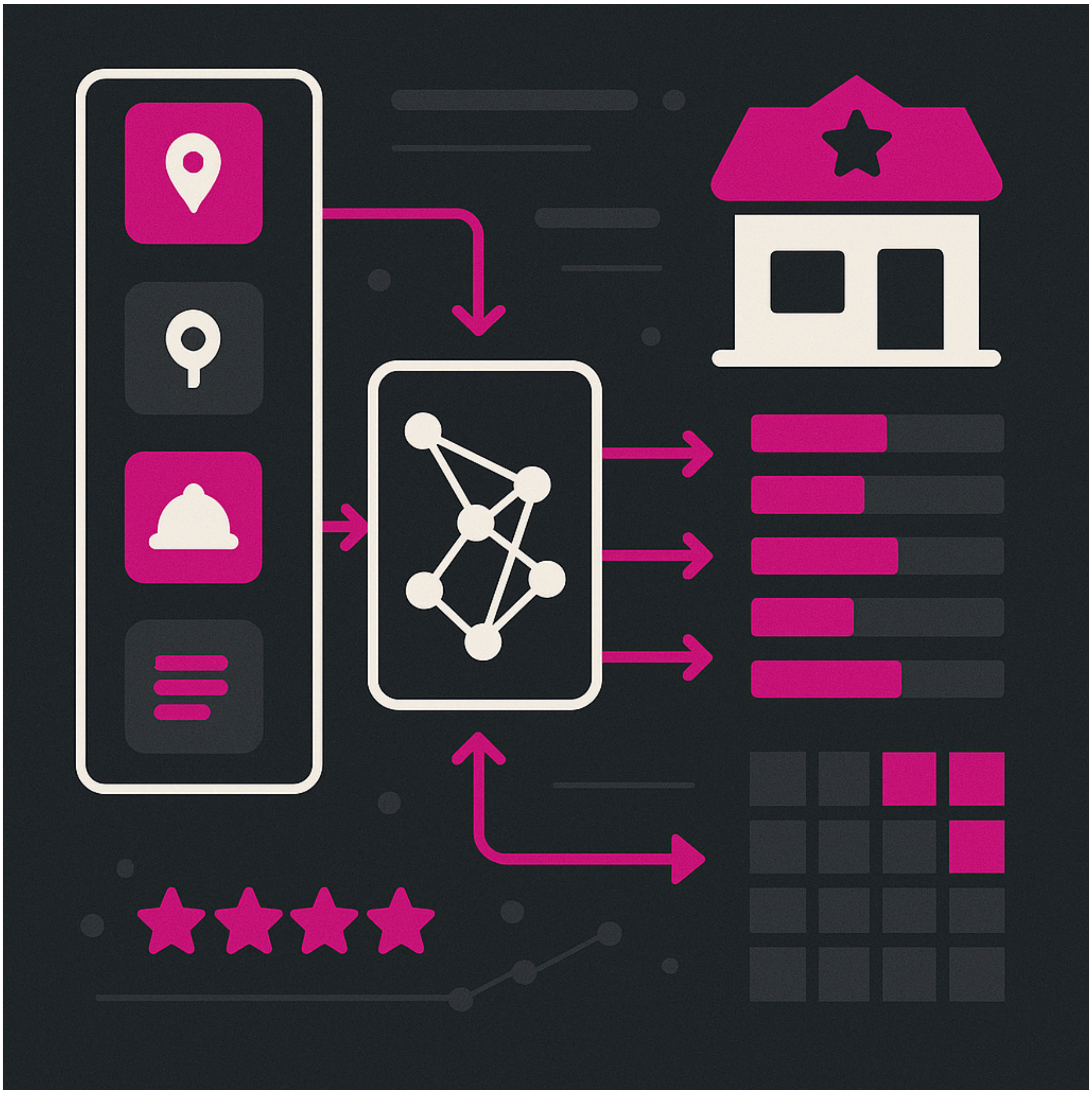
Snappfood | Commercial

Data Analyst Technical Task

Report of Task 2 - Grading Model

Prepared By: **Amirreza Akbari**



Objectives

The purpose of this task is to develop a predictive model that estimates the quality grade of a restaurant based on previously gathered feature data. Grades are defined as integers from 1 to 4, where Grade 1 denotes the best quality and Grade 4 the lowest. The historical dataset used for training includes numeric features describing restaurant performance and behavior, along with manually assigned grades by a previous data analysis team.

Exploratory Data Analysis (EDA)

The dataset contains 399 labeled restaurant records, each with several numerical attributes representing user interactions, marketing activity, and internal metrics. The target variable is `Grade`, an ordinal value ranging from 1 (best) to 4 (worst). The class distribution is highly imbalanced:

Grade	Count	Proportion
1	7	1.8%
2	5	1.3%
3	20	5.0%
4	367	92.0%

This imbalance presented a modeling challenge and influenced preprocessing and evaluation strategies.

Correlation Analysis

Correlation with the target `Grade` revealed strong inverse relationships for several features:

- `Google Sense`: -0.717
- `Survey`: -0.622
- `Search Count`: -0.613
- `Branch Counts`: -0.494
- `Marketing Area`: +0.039

This indicates that higher values of `Google Sense`, `Survey`, and `Search Count` are potentially associated with better (lower-numbered) grades.

Distribution Paterns

Box plots revealed a clear trend for most features:

Grade 1 restaurants consistently have higher values, which gradually decline through to Grade 4.

- `Search Count`, `Google Sense`, `Survey`, and `Branch Counts` all show strong ordinal patterns — with Grade 1 having the highest medians and tighter upper quartiles compared to lower grades.
- `Marketing Area`, while not strongly correlated overall, shows moderate variance across grades. Interestingly, certain marketing area values appear exclusive to specific grade ranges, suggesting location-based tendencies where only high-grade or low-grade restaurants may appear.

Preprocessing

Prior to model development, several preprocessing steps were applied to ensure the data was clean, well-scaled, and suitable for training:

- `ID` **column dropped**: As it contained no predictive value, it was removed from the feature set.
- **Skewness correction**: Four features — `Search Count`, `Survey`, `Google Sense`, and `Branch Counts` — exhibited high positive skewness. A log-transformation using `log1p` was applied to normalize their distributions.
- **Feature scaling**: All numerical features were standardized using `StandardScaler` to ensure uniform contribution to model training.

Modeling

The problem was framed as a multiclass classification task, with attention to the ordinal nature of the target variable (`Grade`). While ordinal regression was not directly implemented, evaluation metrics were chosen to reflect this structure.

The following models were implemented and evaluated:

- **Logistic Regression**: Multinomial logistic regression with regularization.
- **Decision Tree**: Interpretable model capturing non-linear decision boundaries.
- **Random Forest**: Ensemble method combining multiple decision trees for robustness and higher accuracy.

Each model was trained and evaluated using consistent cross-validation and metric strategies.

Imbalance Handling

Given the extreme imbalance in grade distribution — with over 90% of samples belonging to Grade 4 — a class balancing technique was essential.

- **SMOTE** (Synthetic Minority Oversampling Technique) was applied within each fold of cross-validation. This allowed synthetic examples of minority classes (Grades 1–3) to be generated only on the training split, avoiding data leakage.
- **Stratified 5-Fold Cross-Validation** was used to maintain class proportions across folds and ensure reliable evaluation.

Evaluation

Each of the three models was evaluated using Stratified 5-Fold Cross-Validation with SMOTE applied inside each fold to address the class imbalance. The evaluation focused on three key metrics:

- **Accuracy**: Measures overall correct predictions.
- **Macro F1 Score**: Averages F1 scores across all classes, giving equal weight regardless of class size.
- **Quadratic Weighted Kappa (QWK)**: Accounts for ordinal distance between predicted and actual classes, making it ideal for this task.

Results Summary

Model	Accuracy	F1 Score	QWK
Random Forest	0.96	0.84	0.92
Decision Tree	0.93	0.74	0.85
Logistic Regression	0.86	0.66	0.73

Random Forest outperformed the other models across all metrics, especially on QWK, confirming its strength in capturing the ordinal nature of the problem.

Confusion Matrix Insights

- Logistic Regression struggled to differentiate between mid- and low-tier grades (notably misclassifying Grade 3 as Grade 4).
- Random Forest demonstrated superior separation of all grades, with nearly perfect classification of Grades 1 and 4, and relatively strong performance on Grade 3 — the largest minority class.
- Errors in all models tended to shift predictions one grade off, which is less harmful in ordinal contexts but still penalized in QWK scoring.

Feature Importance

After selecting **Random Forest** as the final model, feature importance scores were extracted to identify which input variables had the greatest influence on grade prediction.

The top five features, sorted by descending importance, were:

- `Google Sense`: 0.34
- `Search Count`: 0.30
- `Survey`: 0.16
- `Branch Counts`: 0.14
- `Marketing Area`: 0.06

These results validate the earlier insights from correlation and distribution analysis:

- `Google Sense` and `Search Count` are the most influential predictors, both highly correlated with better grades.
- `Survey` and `Branch Counts` also contribute meaningfully, reflecting customer engagement and business size, respectively.
- `Marketing Area`, while less impactful overall, still plays a supporting role and may interact with other features to guide predictions.

This ranking reinforces that the model is not only accurate but also interpretable — making it more suitable for use by analysts or business stakeholders.

Final Outputs

- `task2.ipynb`

  The complete Jupyter notebook containing EDA, preprocessing, model development, evaluation, and feature analysis.

- `task2_predict.py`

  A standalone script that loads the trained model and scaler to predict grades on new restaurant data provided as a CSV

- `task2_model.pkl`

  The final trained Random Forest model, saved using `joblib` for inference reuse.

- `task2_scaler.pkl`

  The `StandardScaler` instance used during preprocessing, necessary for transforming future data consistently.

Conclusion

A reliable and interpretable model was developed to predict restaurant grades using historical feature data. The final Random Forest model achieved strong performance, including a QWK of 0.916, making it well-suited for the task's ordinal nature.

Key insights included the importance of features like Google Sense, Search Count, and Survey. The use of SMOTE and stratified cross-validation ensured fairness despite class imbalance.

The model, along with the supporting scripts and artifacts, is ready for reuse in real-world applications. While data scarcity in Grades 1 and 2 remains a limitation, the current solution provides a strong foundation for predictive grading tasks.