Predictive Go to Market Strategy Using Machine Learning

Arturo Noguera
May 2025

Berkeley Engineering BerkeleyHaas

What are we solving?

- Out there every customer has plenty of options for every services or infrastructure acquisition , we need to strength our customer´s satisfaction and loyalty.

- At the same time we need to reduce the probability that competitors can take over our clients. Using Data Driven go to Market strategy, we will create product bundles with more attractive prices and promote synergy among different product business units to reduce production and marketing costs.

- Using different prediction models, we will create and deploy a tool that can predict if a new customer is candidate to accept a direct offer of product bundles from two main categories in our product portfolio.

Why does it matter?

AC4MEX INC, a large IT firm, recently merged with one of their competitors, each original firm is now an independent Division of the new corporative group.

Even when some of their products overlap, they keep complementary product portfolios, and look forward to approach the combined customer base with product bundles from both division´s catalog.

Current challenges

- Inventory control & Supply-chain is not consolidated between two main Divisions
- Marketing efforts and budgets are independent
- The opportunity to reach customers from each other Division is diluted

What is the business goal?

- Our challenge is get to know each customer´s "buying profile" and target specific customers with higher probability to buy bundles from a broader portfolio vs individual products.

- **By performing different ML models, we aimed to improve at predictively optimizing following:**

- Operational efficiency
- Inventory optimization
- Reduce Marketing Costs
- Improve customers satisfaction
- Create synergy between product divisions

How did we achieve this?

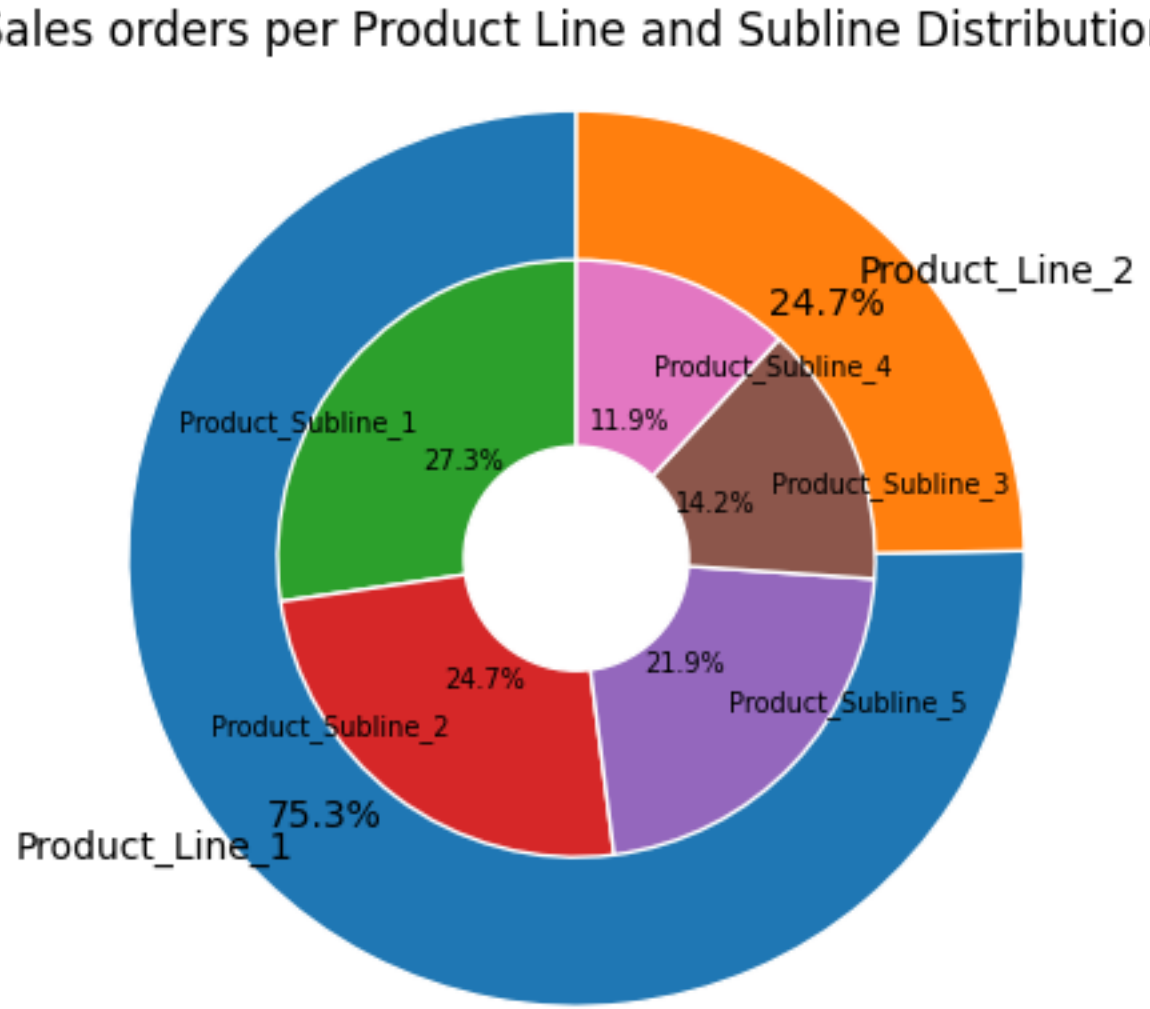
- In order to predict which customers have more probability to buy items from any category belonging to Product_Line_1 , given that previously they bought items from any category belonging to Product_Line_2 we will analyze and model a Dataset including +15000 Sales orders from final customers
- The Data was obtained from ACMEX INC Annual Bookings report. It is synthetic in nature and anonymized to obfuscate identifiable information.

To deploy our prediction tool , we used the following AI/ML tools:

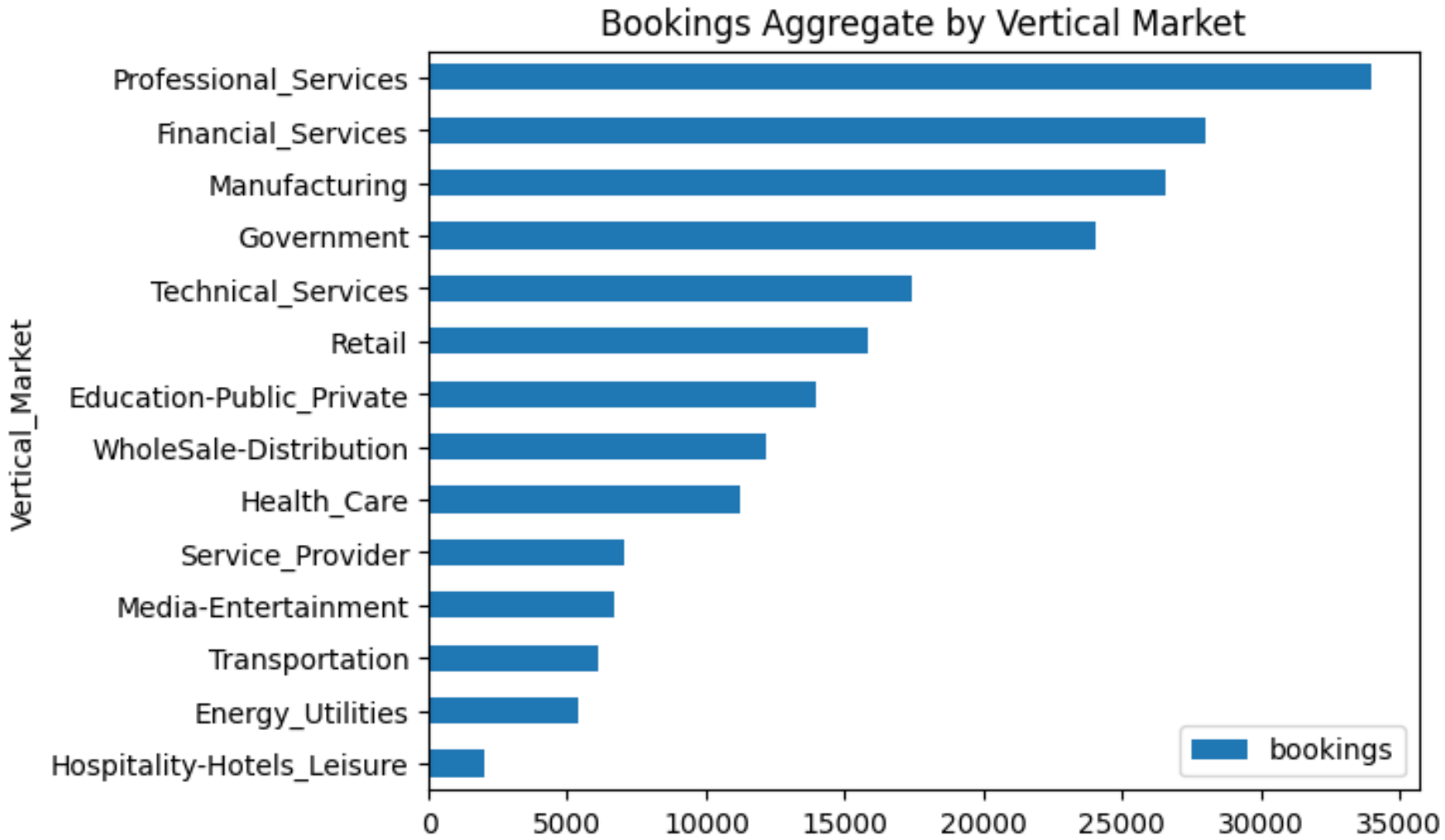
- **Why a Classifier Ensemble?**
- A classifier ensemble combines multiple classifiers to enhance classification accuracy and generalization. It's a technique where diverse individual classifiers are trained, and their predictions are combined using various methods like voting or averaging.

- **Why Oversampling ?**
- Imbalanced datasets impact the performance of the machine learning models and the Synthetic Minority Over-sampling Technique (SMOTE) addresses the class imbalance problem by generating synthetic samples for the minority class

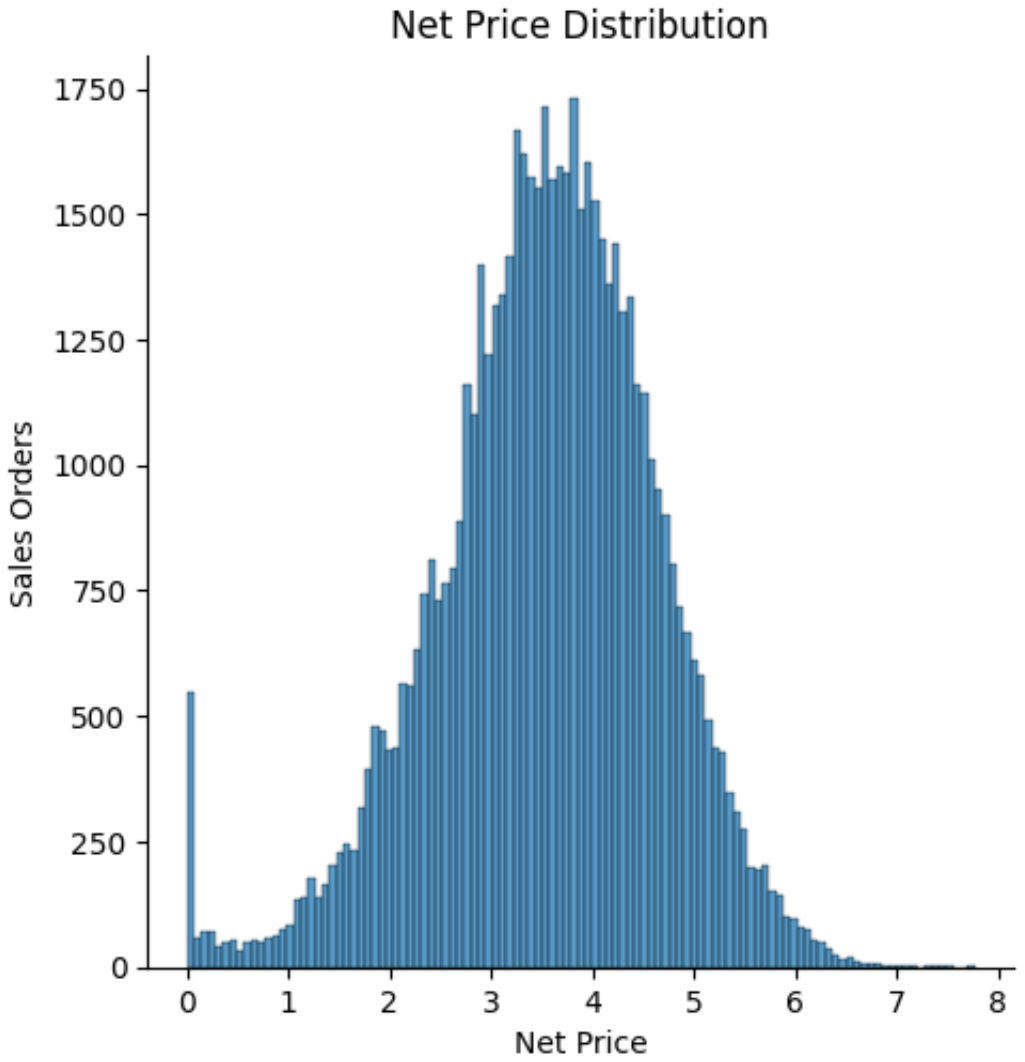
Data Insight



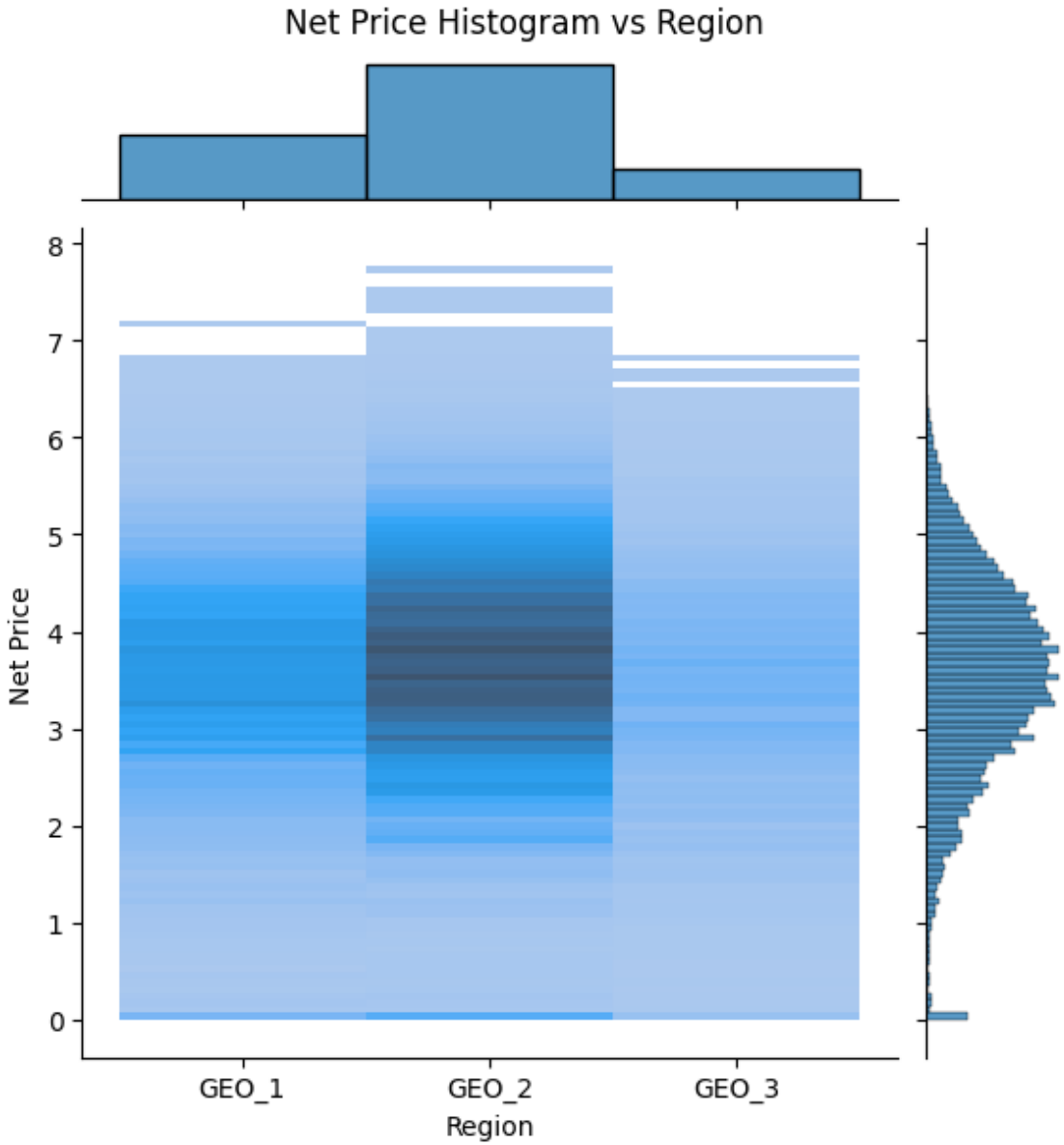
Data Insight



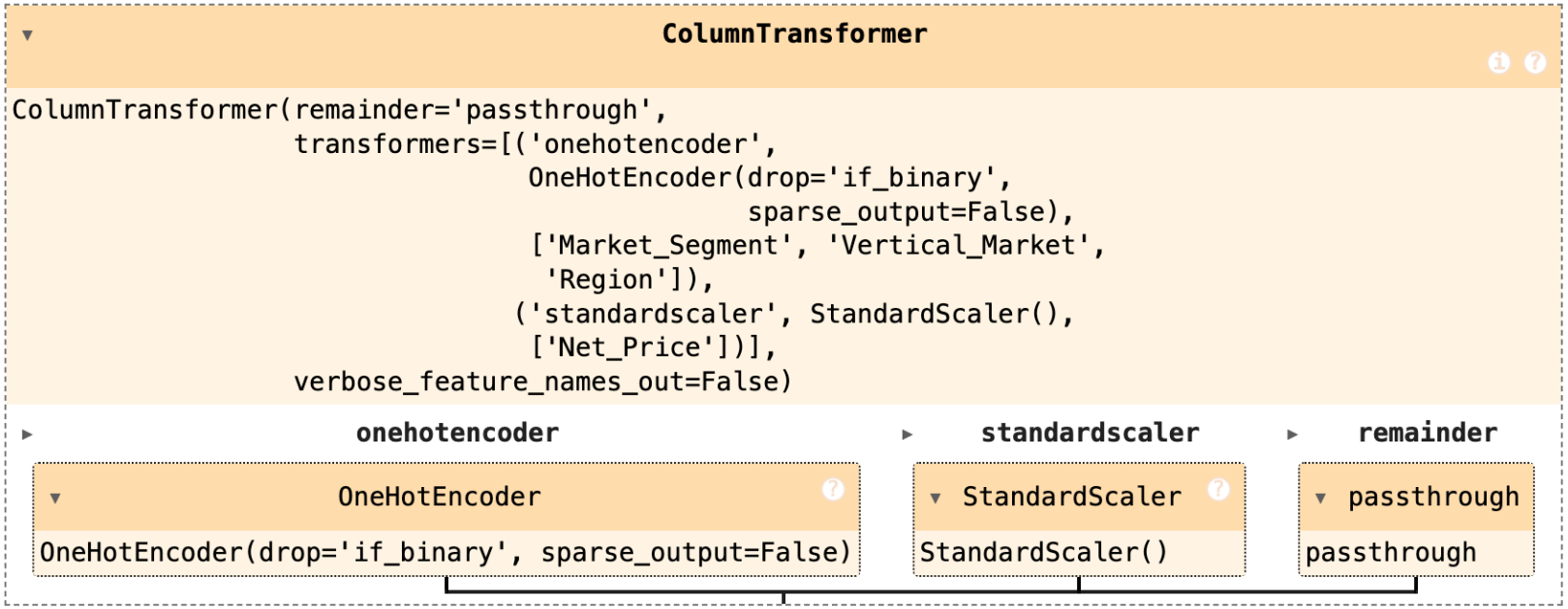
Data Insight



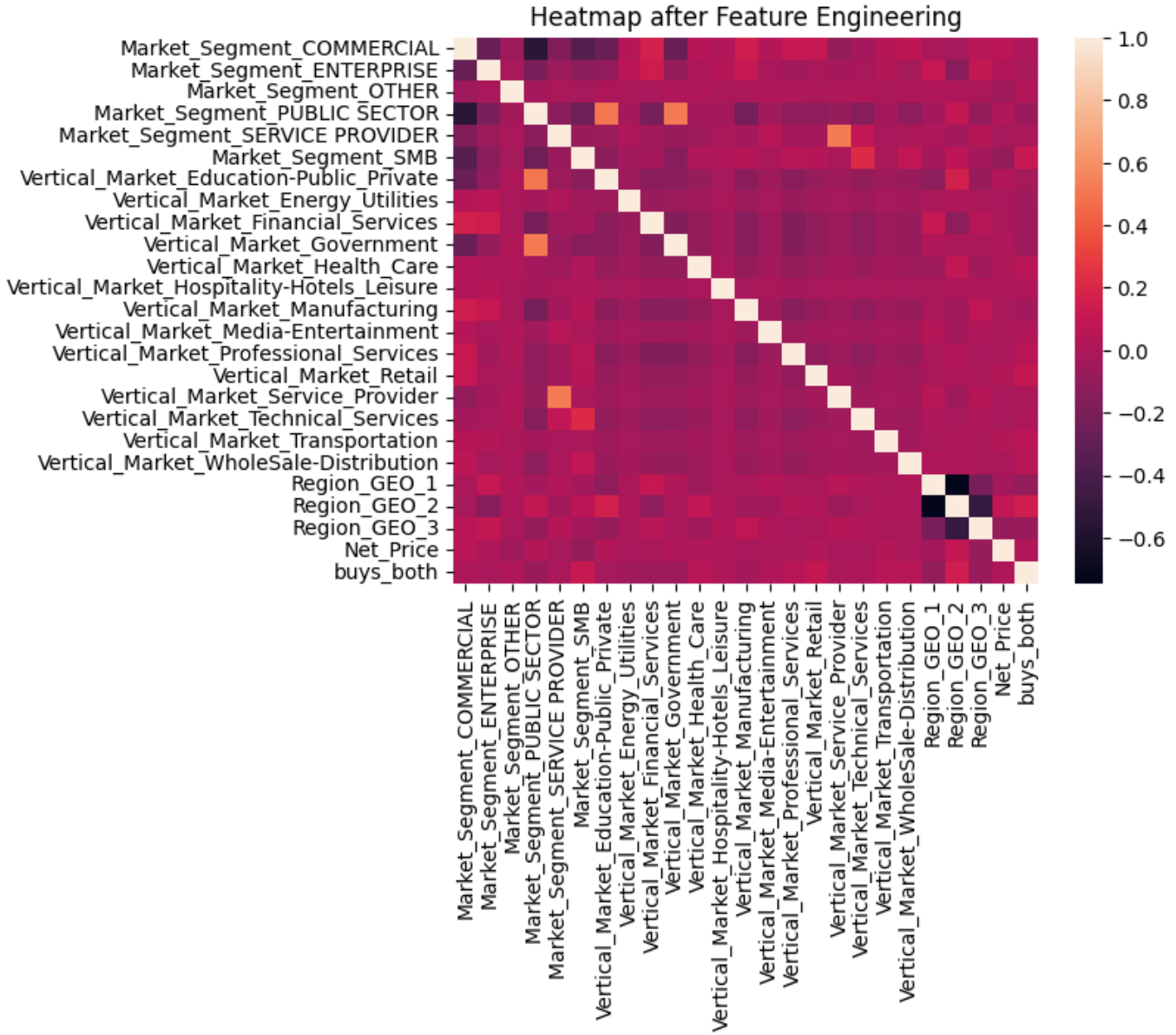
Data Insight



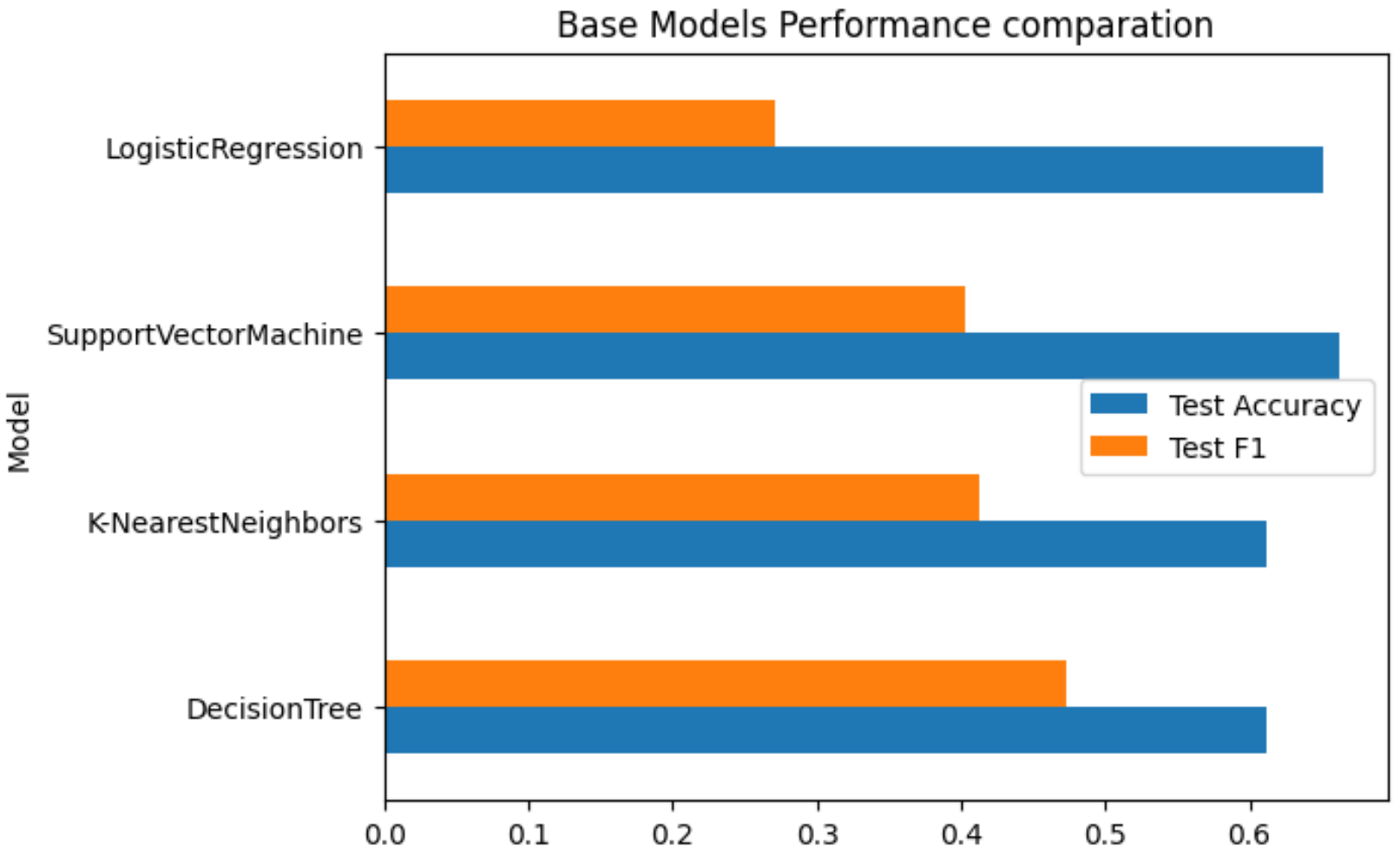
How we engineered model features ?



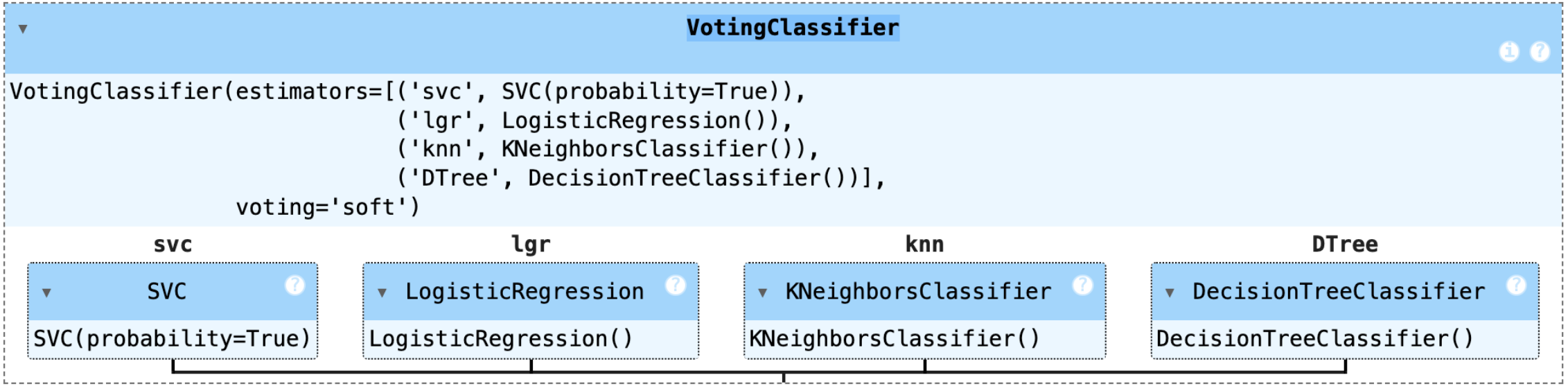
Correlation Analysis



Base Models Training and Scoring



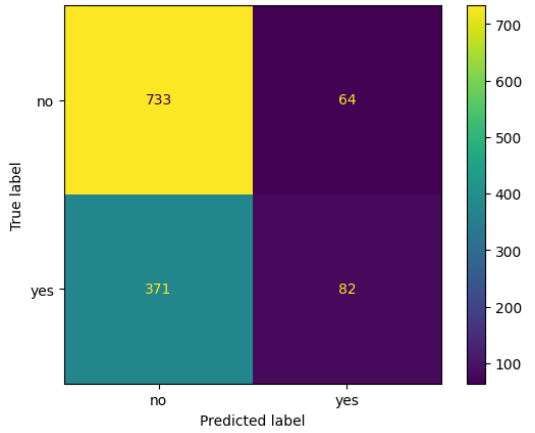
Base Models Voting Ensemble and Scores

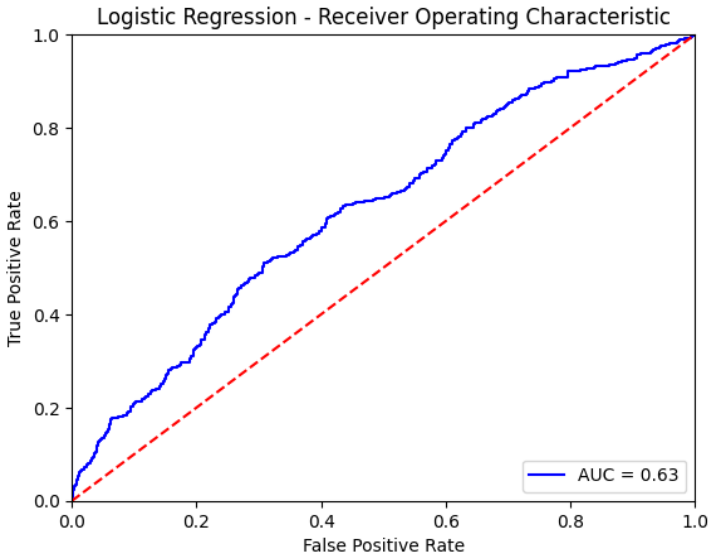


Model	F1	Accuracy	Precision	Recall
SOFT Voter_Base Models	0.424	0.626	0.479	0.380

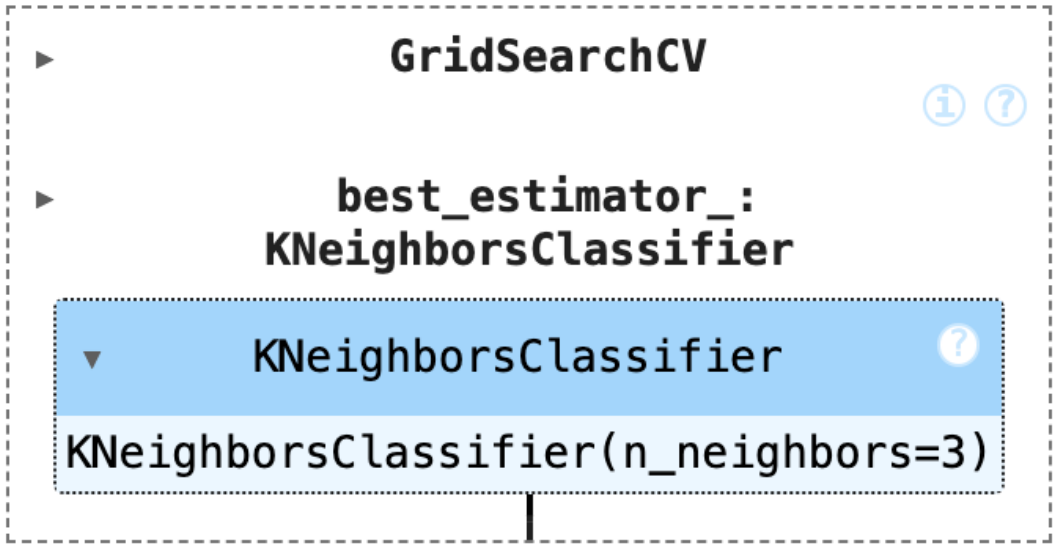
Confusion Matrix for Logistic Regression with Grid Search

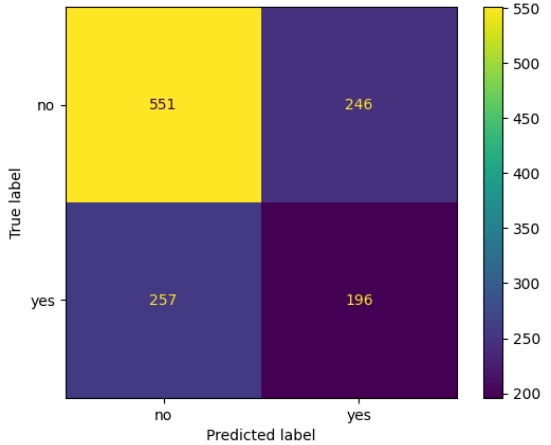
```
GridSearchCV
best_estimator_: LogisticRegression
LogisticRegression
LogisticRegression(fit_intercept=False, solver='liblinear')
```

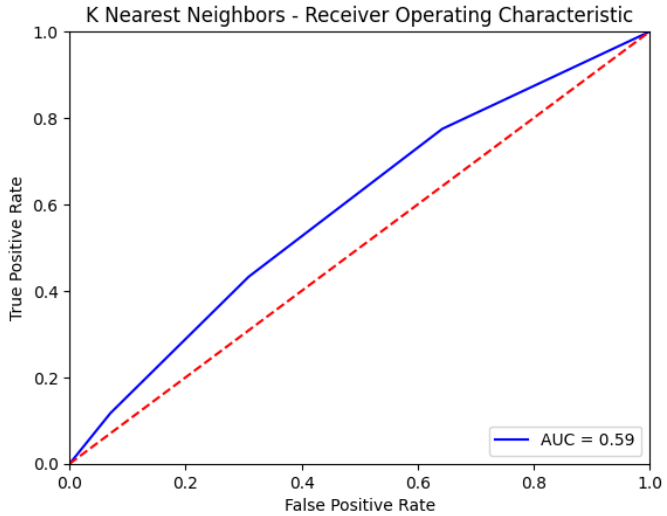




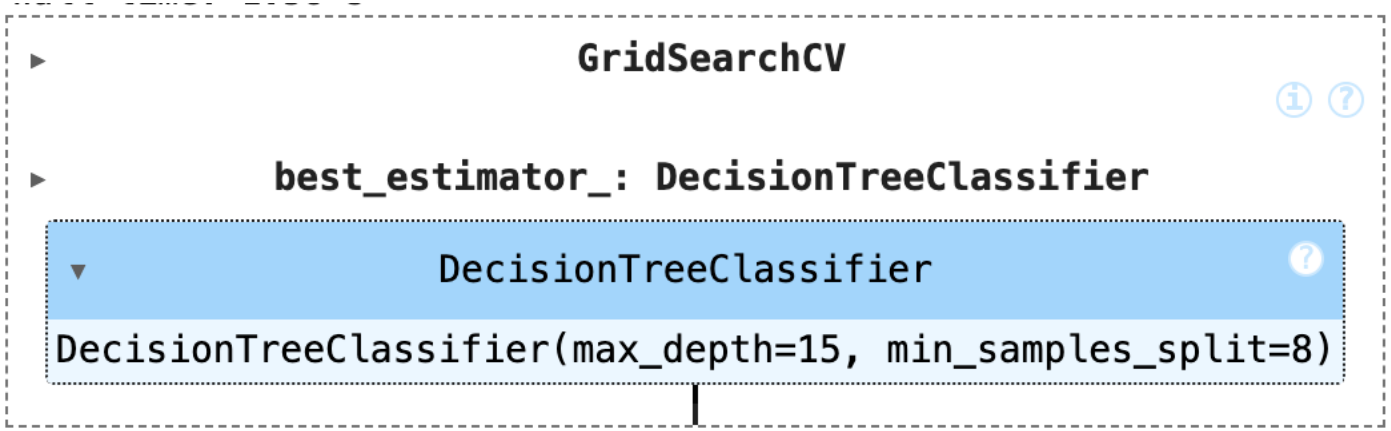
Confusion Matrix for **K Nearest Neighbors** with Grid Search

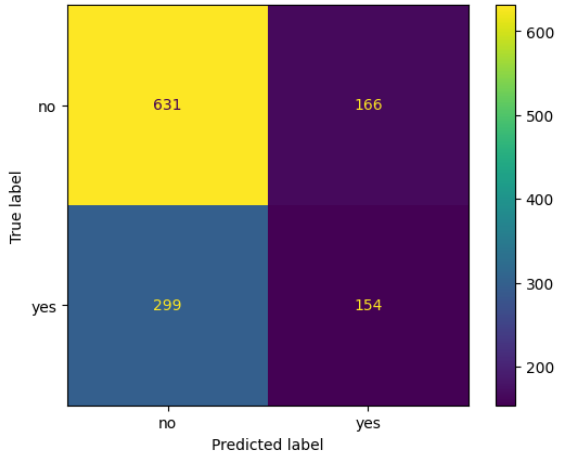


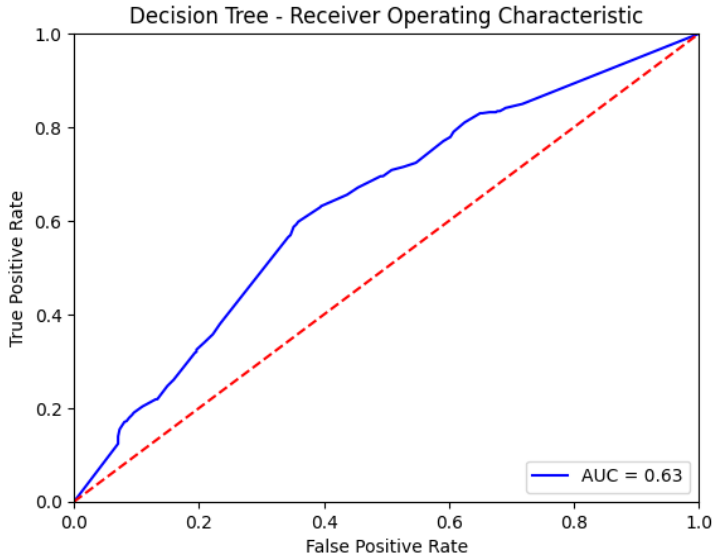




Confusion Matrix for **Decision Tree with** Grid Search

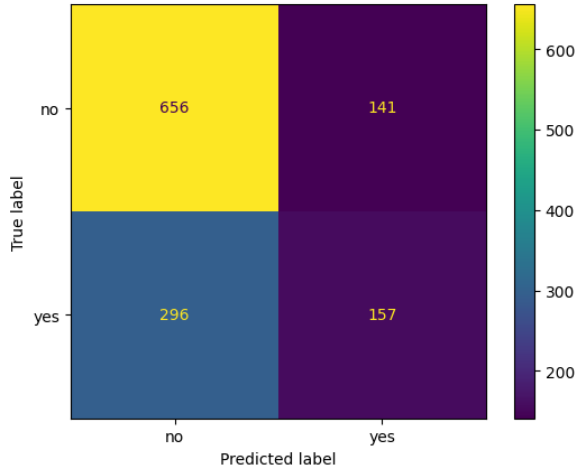


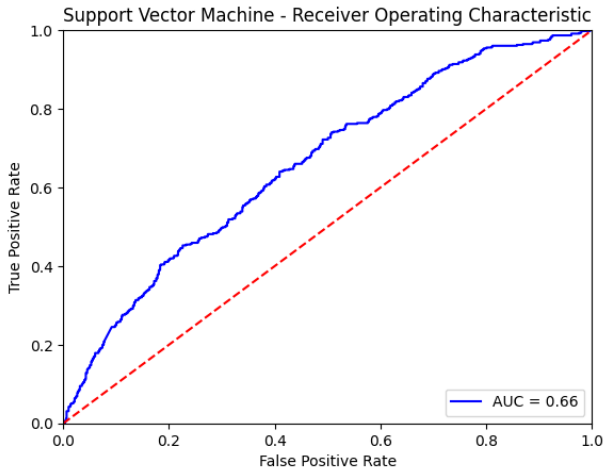




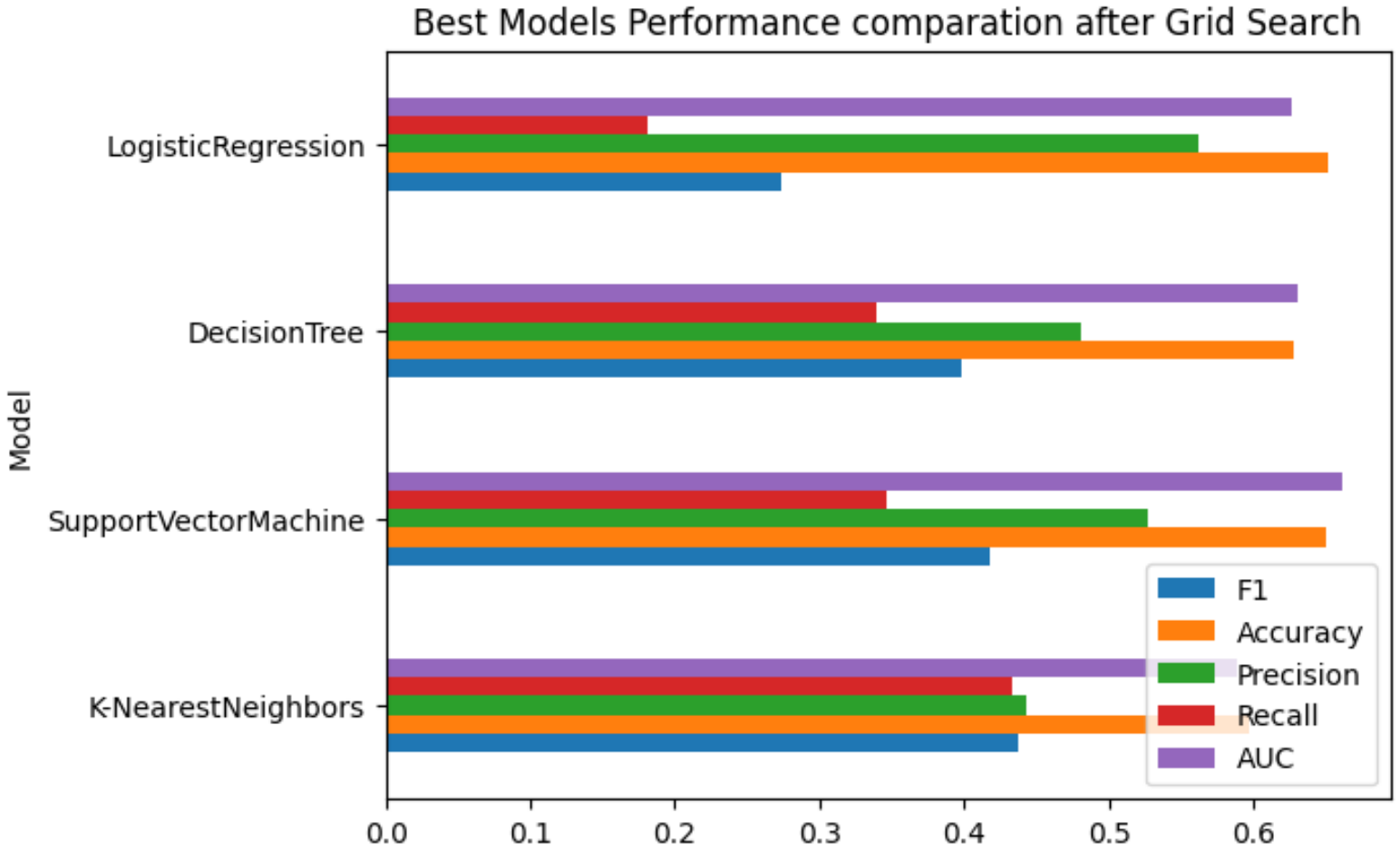
Confusion Matrix for Support Vector Machine **with** Grid Search

```
GridSearchCV
best_estimator_: SVC
SVC
SVC(C=10, degree=2, probability=True)
```

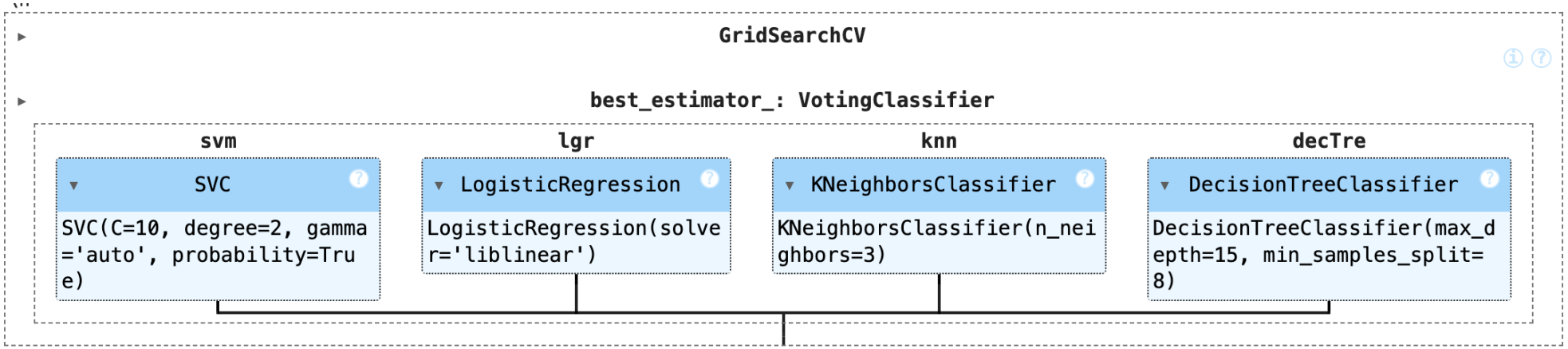




Model´s Performance after tuning with Grid Search



Tunned Models Voting Ensemble and Scores



Model	F1	Accuracy	Precision	Recall	AUC
SOFT Voter_Grid_Search	0.376	0.660	0.561	0.283	0.655

Synthetic Minority Oversampling Technique (SMOTE)

- Due to the imbalanced nature of the target Variable in the initial Data Set, we will test another technique call SMOTE.

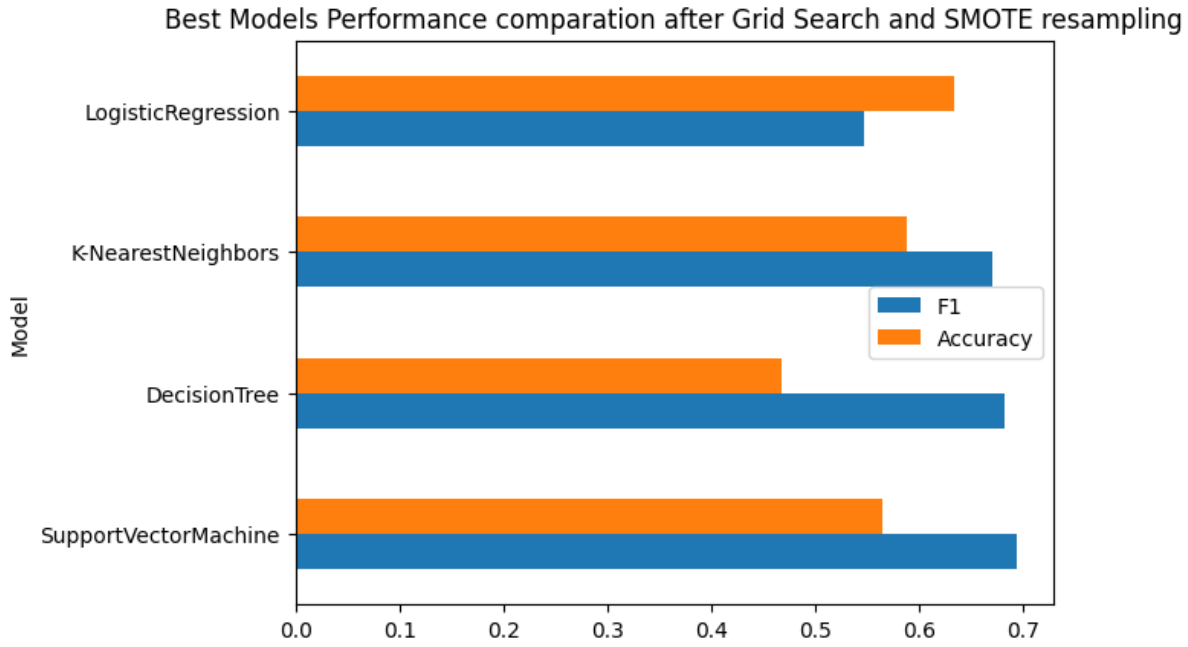
- Synthetic Minority Oversampling Technique (SMOTE) is a popular method to over-sample minority class.

- SMOTE should be applied only to the training data, after it has been separated from the testing data.

- This prevents data leakage and provides a more reliable evaluation of your model's performance on unseen data

What was our outcome?

Model	F1	Accuracy
SupportVectorMachine	0.694	0.565
DecisionTree	0.682	0.468
K-NearestNeighbors	0.671	0.588
LogisticRegression	0.547	0.634



Summary

After executing oversampling on the minority class using SMOT over all four models with GridSearch, the F1 performance increased significatively for all four models

- SupportVectorMachine reported 0.695 F1 score,
- DecisionTrees reported 0.690 F1 score,

these figures are superior to the Accuracy figures reported with the same tuned models.

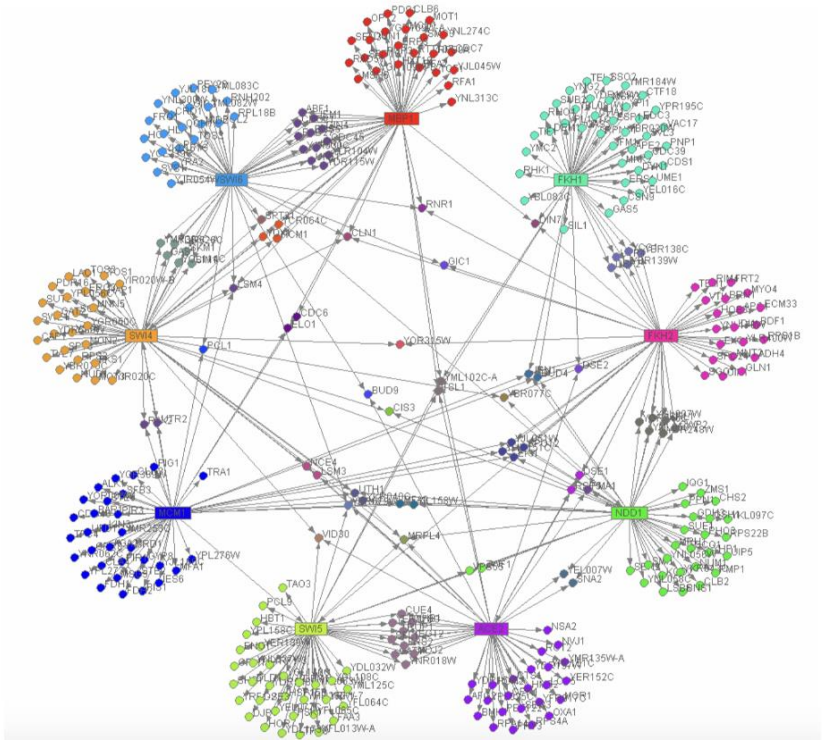
Given the imbalanced distribution from our Target Variable, we will deploy resampling using SMOTE, and the following tuned models

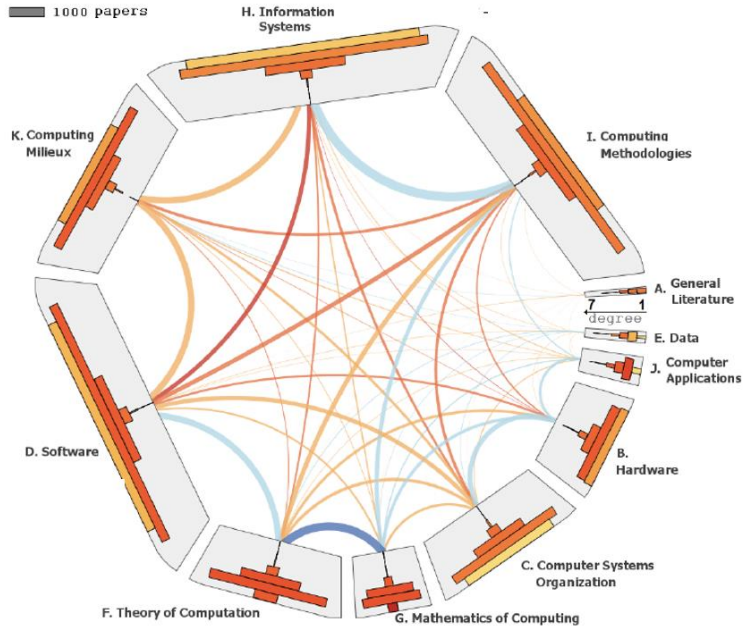
- SupportVectorMachine
- DecisionTrees

Next Steps and recommendations

- The correlation analysis after feature engineering, shows poor positive correlation of most variables vs our target variable. We may need to change the way the original data was extracted and add additional features that may lead to significant correlation to target variable.
- Including dates in each Sales Order, could lead to an additional feature that models the procurement sequence from products of interest for each customer.
- Resampling methods increased marginally the accuracy, so in this scenario SMOT is not a final solution to the imbalance problem. It's important to understand that imbalanced data is a problem only if the minority class in the training set is not representative enough and/or the features are not good enough indicators for the label.

Q&A





Thank You

Special Thanks to Savio Saldanha for the mentoring and advice