

# Predicting Loan Approval:

## Binary Classification of Financial Risk Dataset

Oliver Butterworth-Bakhshi

# Executive Summary

## Summary of findings

- Objective: Use ML models to predict loan classification from a Financial Risk dataset.
- Use case: ML model can reduce the number of false positives (based on specificity), saving the bank millions of pounds in customer mortgage defaults that would otherwise be avoided.
- Data Summary:
  - (Before processing): 20000 rows, 36 columns. (After cleaning and pre-processing): 5452 rows, 42 columns.
  - Target variable: y, Loan Approval, 1 for approved, 0 for rejected.
  - Key preprocessing steps: Removal of outliers, Drop irrelevant columns, Train-test split, Transform and scale categorical columns using Transformer, get.dummies() and StandardScaler.
- Modelling and results:
  - Models used: kNN, SVC, Decision Tree, Bagging, Pasting, Random Forest, XGBoost, lightgbm.
  - Evaluation metrics: Precision, Accuracy, Recall, Specificity, with a focus on precision and specificity to maximise business benefit financially.
  - Findings: Best performing model overall = SVC. All models (apart from kNN) had over 0.96 in all metrics scored, very good fitting of data to model.
  - Key Insights: EDA highlighted that higher educated applicants were more likely to be successful. Feature importance and SHAP analysis highlighted key predictors of loan approval as RiskScore, Annual Income and Total DTI.
- Recommendations and further study:
  - Recommendations:
    - Use best predictive SVC model to implement classification of loan applications.
  - Further study:
    - Use local SHAP analysis of best performing model to further validate conclusions.
    - Either improve the performance of feature-compatible models, or find a way to show feature analysis for the best performing SVC model.
    - Estimate business impact for a single bank: request and use a company dataset in a similar manner to this study.

# Introduction

## Background, Motivation, why data-driven approach?

- “Lenders view loan approval from a risk management standpoint”. <https://fastercapital.com/topics/understanding-the-importance-of-loan-approval.html>
- By carefully assessing each applicant's creditworthiness, lenders minimise the chances of default.
- When a customer defaults on their loan, Banks typically lose 30-50% of the total loan amount. Average mortgage is 200K in the UK. Default rate is 5%.
- Predictive models could save banks millions. Hence banking sector needs a data-driven analysis to better assess customer creditworthiness.

# Method

## Data pre-processing

- Datasets from Kaggle: <https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval>, <https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data>. I chose to focus solely on the Financial Risk dataset as it had a greater depth of columns to choose from.
- Dropped columns: UtilityBillsPaymentHistory, BankruptcyHistory, PreviousLoanDefaults - these columns were irrelevant to my analysis.
- Dropped column - MonthlyIncome. This column was very similar in correlation to Annual Income so to get greater contrast in the model, I removed MonthlyIncome rather than have both columns.
- Outliers - errors I saw in the dataset were removed:
  - Experience > 50 years
  - Loan duration > 40 years
  - Credit score < 430
  - Savings >> 100k and current account balances >> 50k.
- Final Dataset size: x\_train = (5452 rows, 42 columns), x\_test = (1364 rows, 42 columns).
- Used get\_dummies on categorical columns, transformed data with OneHotEncoder and StandardScaler.

# Method (cont.)

## Choice of ML models

- I used a broad range of ML models for classification analysis.
- RandomForest, XGBoost, kNN, SVC, Pasting, Bagging, lightgbm, DecisionTree classifiers.
- Approach: fit all the models and compare performance.
- I used GridSearchCV and Bayesian Optimisation to find the optimal combination of hyper-parameters for each model.
- Hyper-parameter scoring prioritised test precision, as the cost of false positives is high (banks who approve too many risky loans can lose a lot of money if the customer defaults).

# EDA insights

## Summary: Part 1

- Loan approval rate = 27.58%. 7 in 10 loans are not approved.
- Education is a key factor in getting approval for a loan:
  - The number of high school graduates is 33% for rejection, and 18% for approval.
  - The proportion of approved applicants increases with the level of higher education. Doctorates are most likely to be approved, followed by Masters and then Bachelor applicants.
- Annual Income: high annual income correlates with loan approval.
  - All applicants with >200k yearly salary were successful in getting a loan.
  - >100k, 85.63% were approved for a loan.
  - <100k, 80.68% were not approved for a loan.

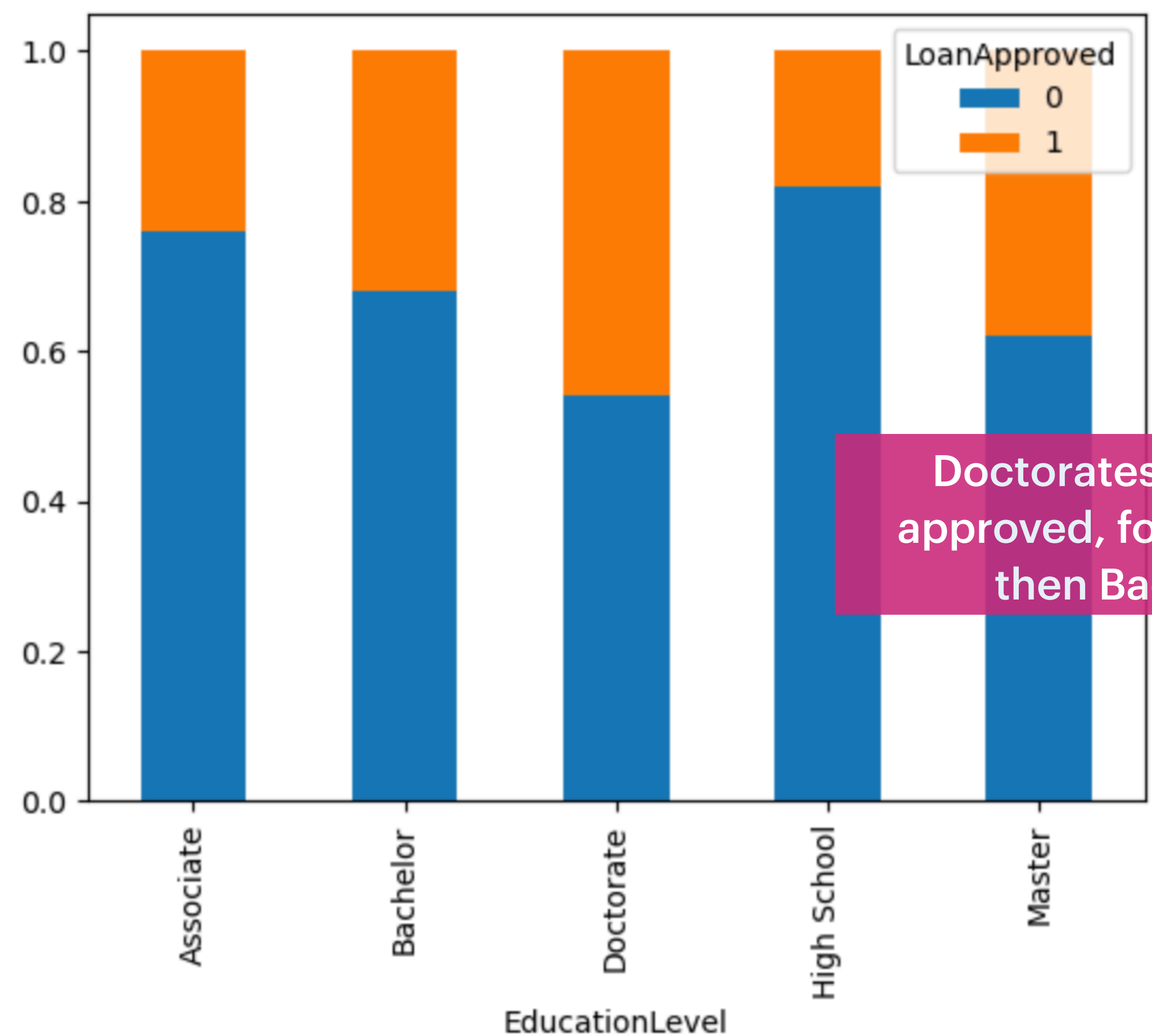
# EDA insights

## Summary: Part 2

- Self employed more likely to have their loan approved than employed/unemployed.
- Loan amount/Monthly Loan Payment: Applicants asking for a higher loan amount and who would be paying off a large amount of debt monthly on a proposed loan are more likely to be rejected, as the lender views them as a high risk investment. Below a proposed monthly loan payment of 1k, 38% of applicants were successful in getting a loan. Above 2k: 90.10% of applicants were rejected for a loan.
- TotalDTI: Low DTI combined with high annual income leads to loan approval.  $\text{mean}|\text{Total DTI}| = 0.139$  with an annual income of 115k+, 90% of loans were approved.

# EDA insights

## Graphical Summary: Part 1



Doctorates are most likely to be approved, followed by Masters and then Bachelor applicants.

LoanApproved		proportion
0		72.42%
1		27.58%

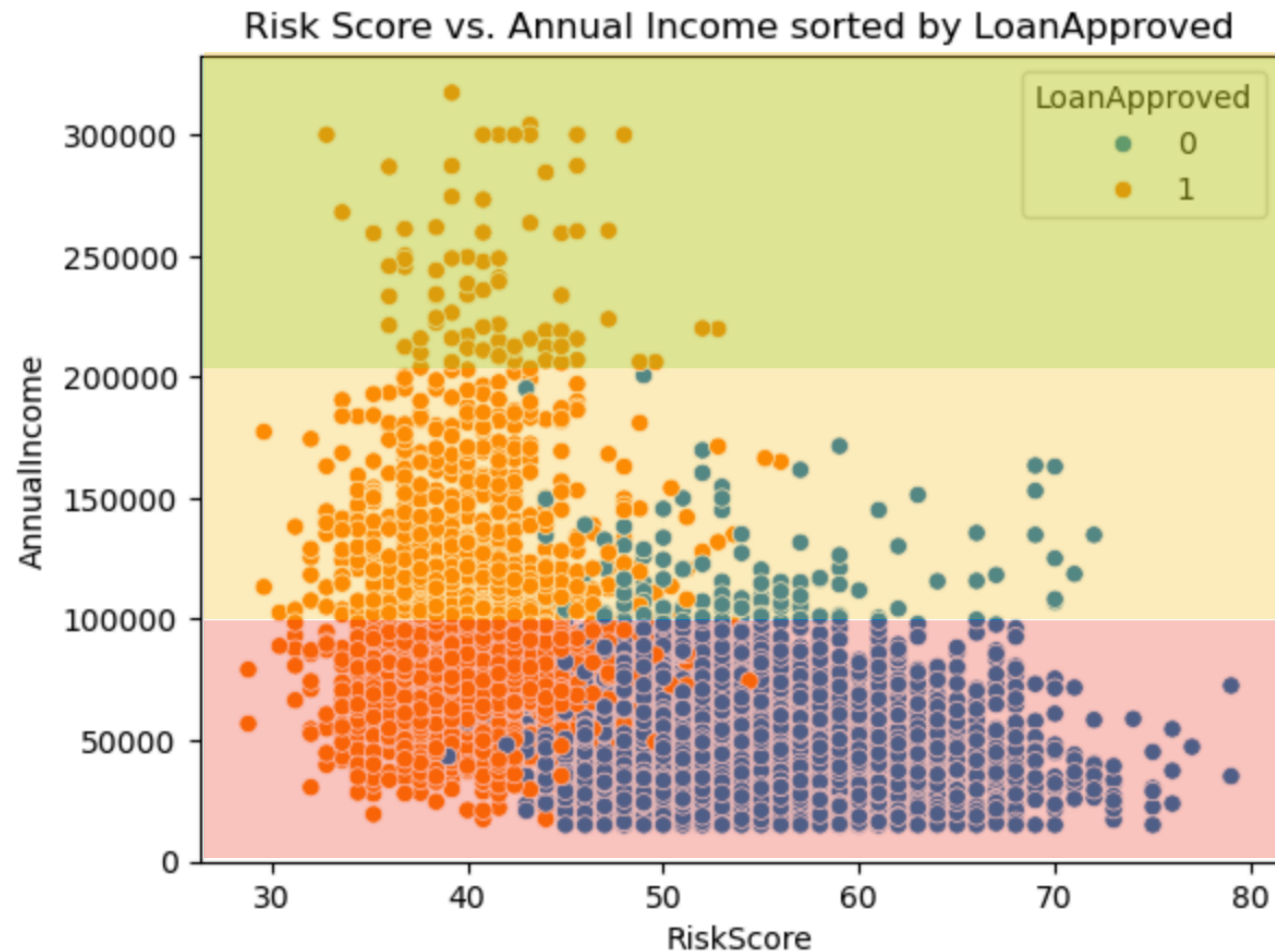
Loan approval rate = 27.58%

LoanApproved		EducationLevel	proportion
0		High School	33.55%
		Bachelor	29.11%
		Associate	20.77%
		Master	13.07%
		Doctorate	3.50%
1		Bachelor	35.21%
		Master	21.22%
		High School	18.72%
		Associate	16.86%
		Doctorate	7.98%



# EDA insights

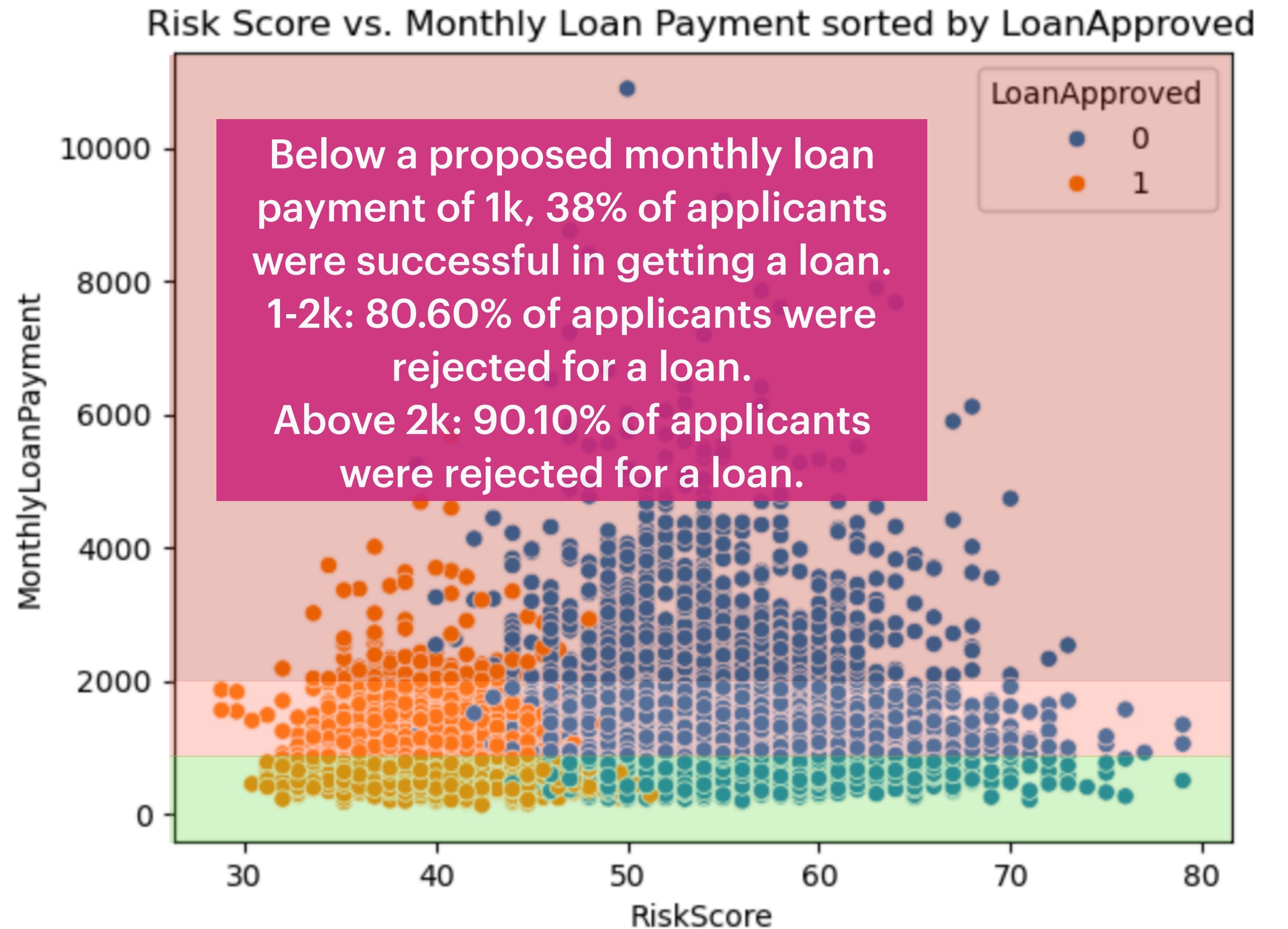
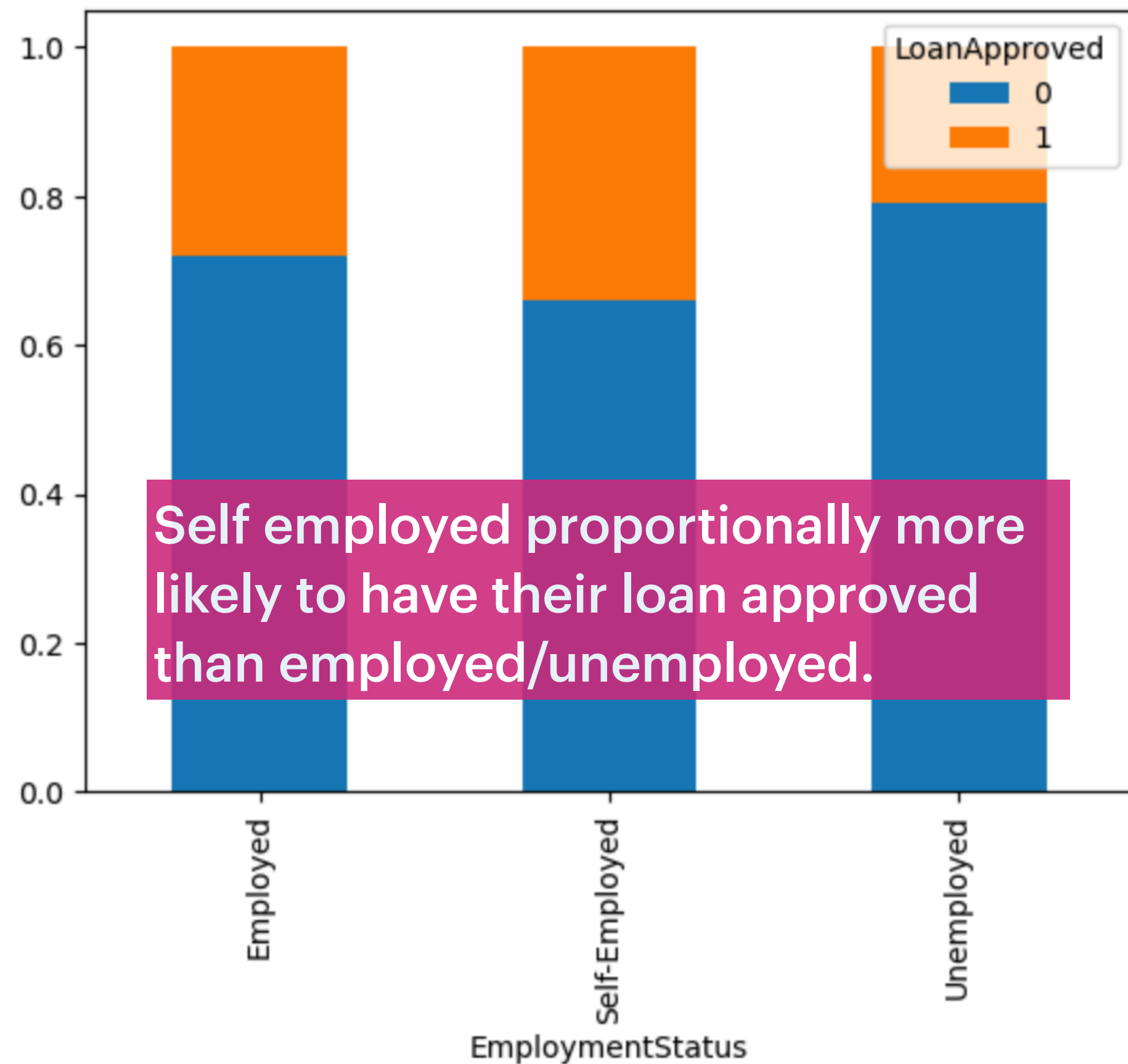
## Graphical Summary: Part 2



All applicants with >200k yearly salary were successful in getting a loan.  
>100k, 85.63% were approved for a loan.  
<100k, 80.68% were not approved for a loan.

# EDA insights

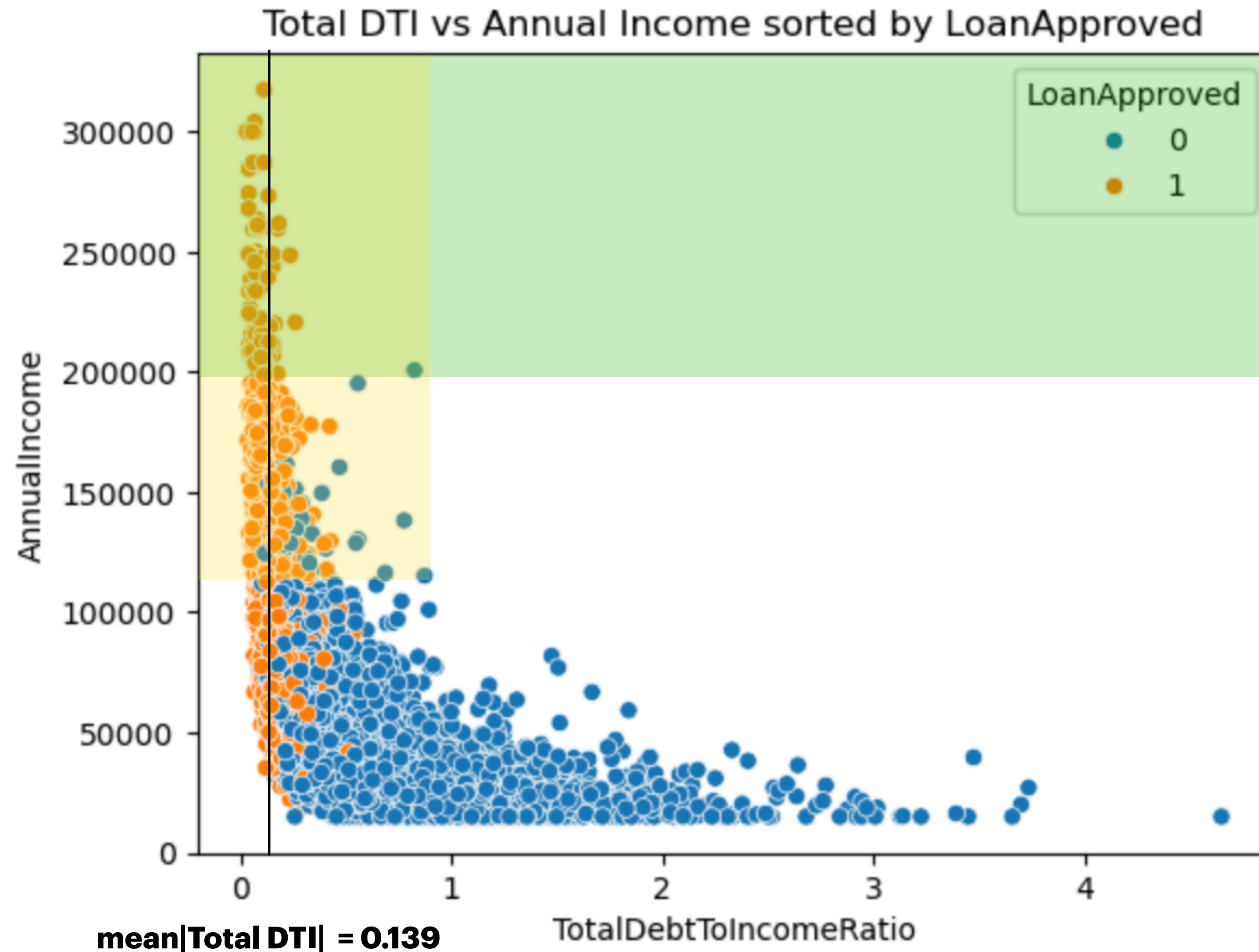
## Graphical Summary: Part 3





# EDA insights

## Graphical Summary: Part 4



All applicants with >200k yearly salary were successful in getting a loan.

mean|Total DTI| = 0.139 with an annual income of 115k+, 90% of loans were approved.

# Modelling Results

## Performance Evaluation: Table

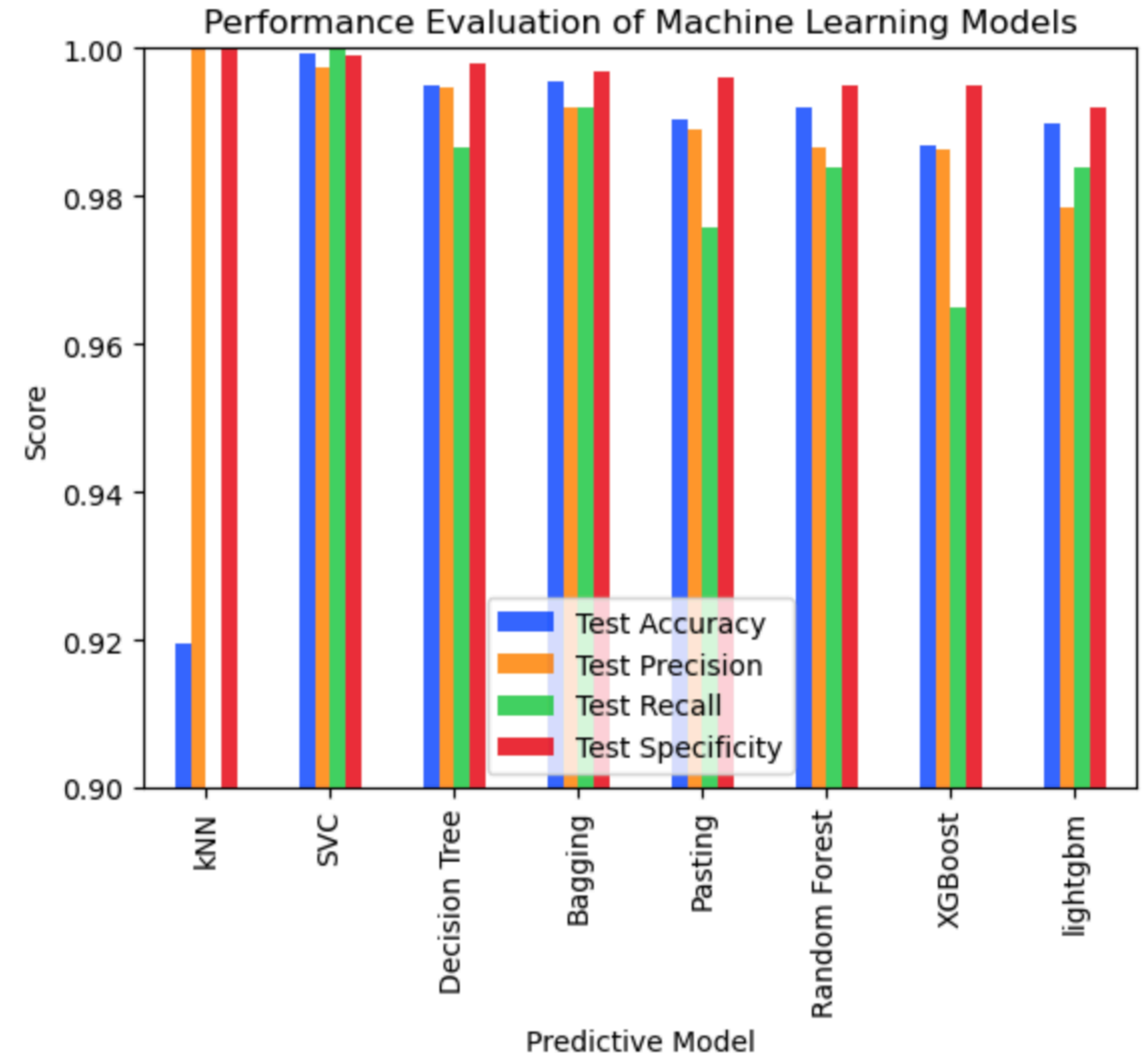
- Models sorted by test precision.
- Most precise model: kNN
- Most accurate model: SVC
- Best performing model overall: SVC

	Model	Test Accuracy	Test Precision	Test Recall	Test Specificity
0	kNN	0.919355	1.000000	0.703504	1.000000
1	SVC	0.999267	0.997312	1.000000	0.998993
7	lightgbm	0.994868	0.994565	0.986523	0.997986
6	XGBoost	0.995601	0.991914	0.991914	0.996979
4	Pasting	0.990469	0.989071	0.975741	0.995972
3	Bagging	0.991935	0.986486	0.983827	0.994965
2	Decision Tree	0.986804	0.986226	0.964960	0.994965
5	Random Forest	0.989736	0.978552	0.983827	0.991944

# Modelling Results

## Performance Evaluation: Graphical Comparison

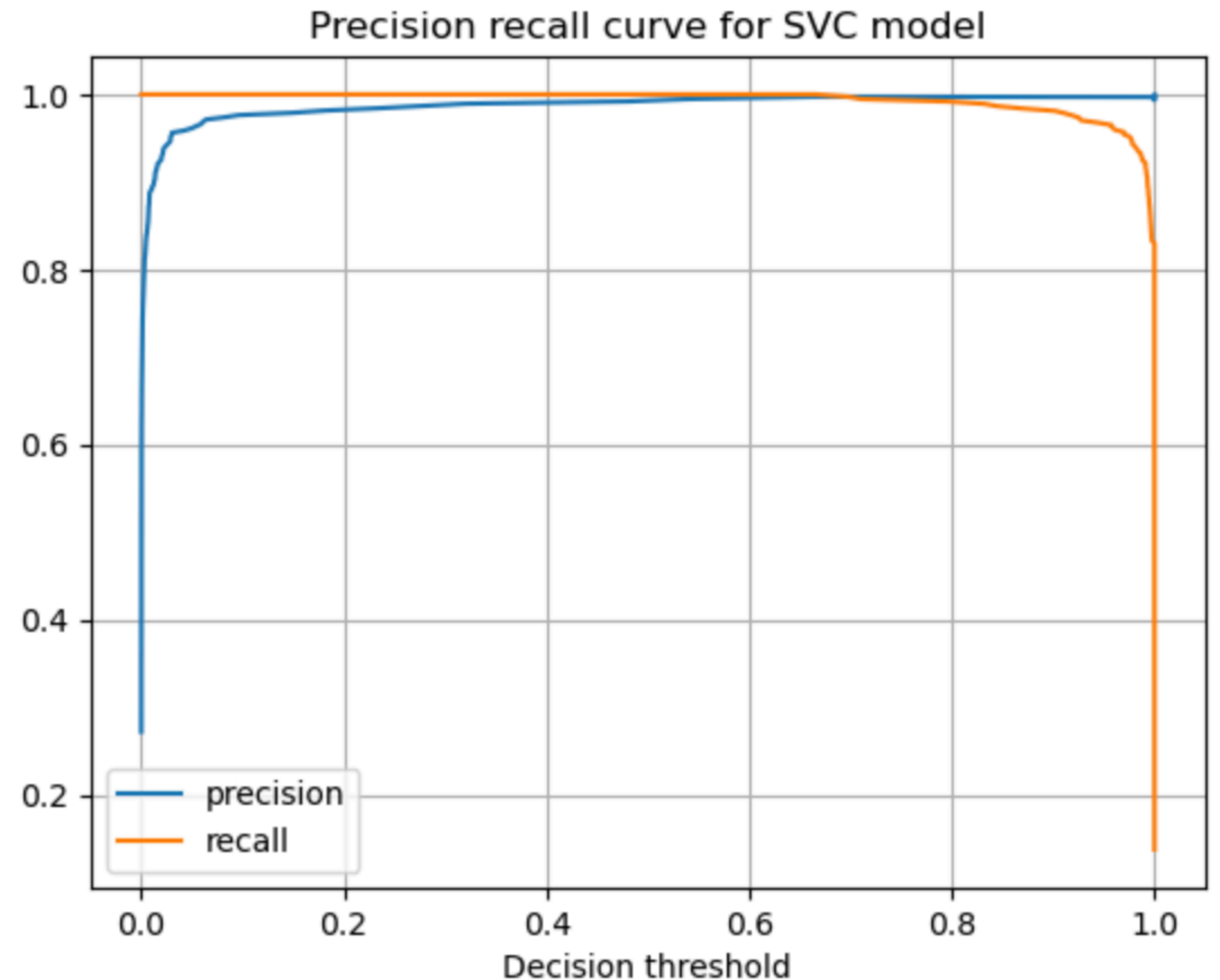
- kNN precise but less accurate.
- SVC strong all round performance
- Apart from kNN, none of the models dropped below 0.96 - very well fitted models and strong classification performance.
- Validates feature importances in next slides



# Modelling Results

## PR curve for overall best performing SVC model

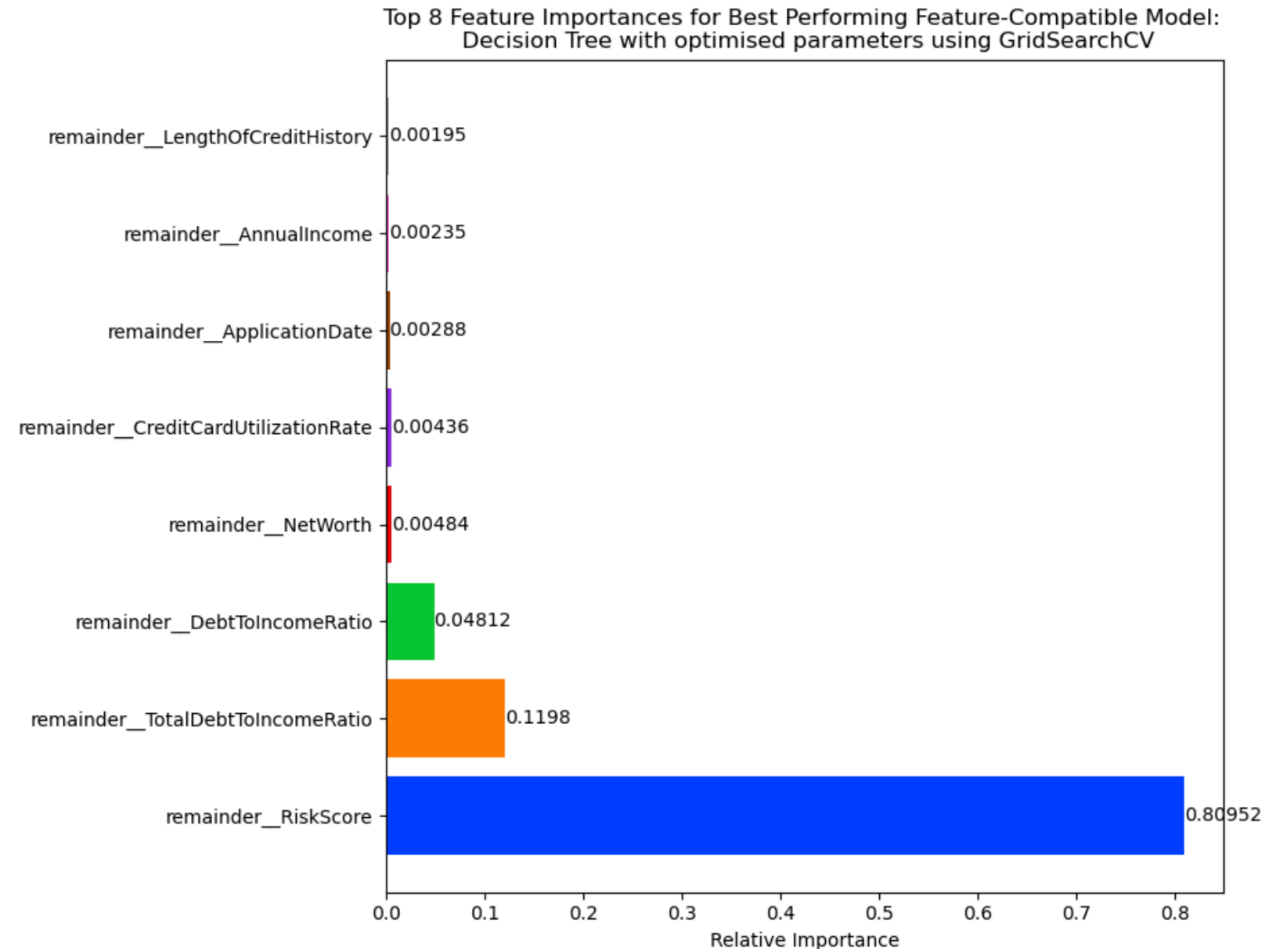
- Very sharply defined decision threshold.
- AUC (Area under curve) is approximately 1 (or very close to 1).
- This suggests an optimal, near-perfect fit.
- Performance in classification is very high, near 100%.



# Modelling Results

## Feature importances: Decision Tree model

- Risk score overwhelming indicator of whether loan is approved.
- Other major factors include TotalDTI, DTI, Net Worth
- Lower income applicants with large debt payments have higher risk scores and are more likely to be rejected for a loan.

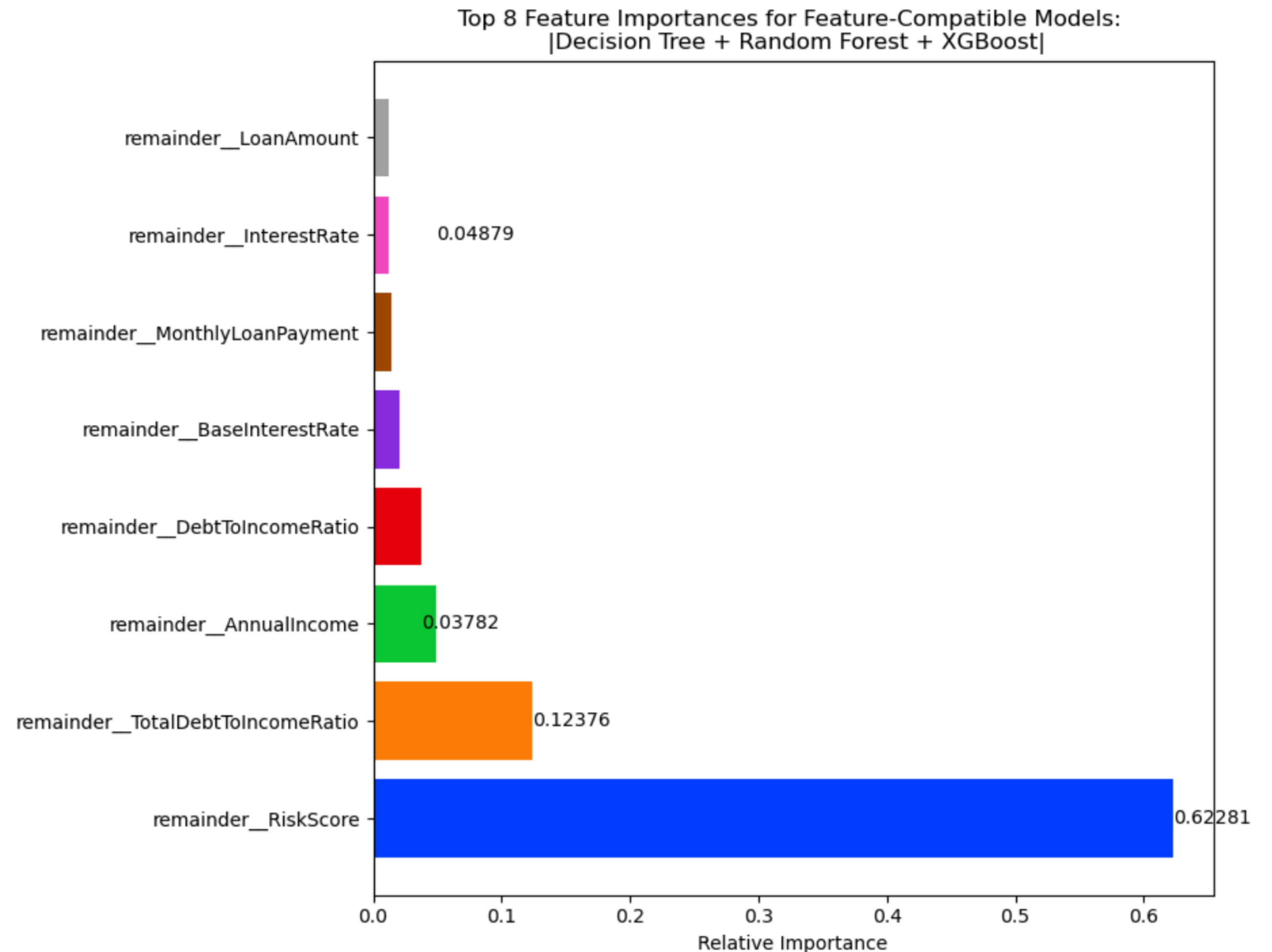




# Modelling Results

## Feature importances: |DT + RF + XGBoost|

- Risk score biggest factor again.
- TotalDTI, Annual Income and DTI key factors in predicting loan approval.
- This matches the conclusions found in EDA: Higher incomes, lower debts, lower risk score -> loan more likely to be approved.

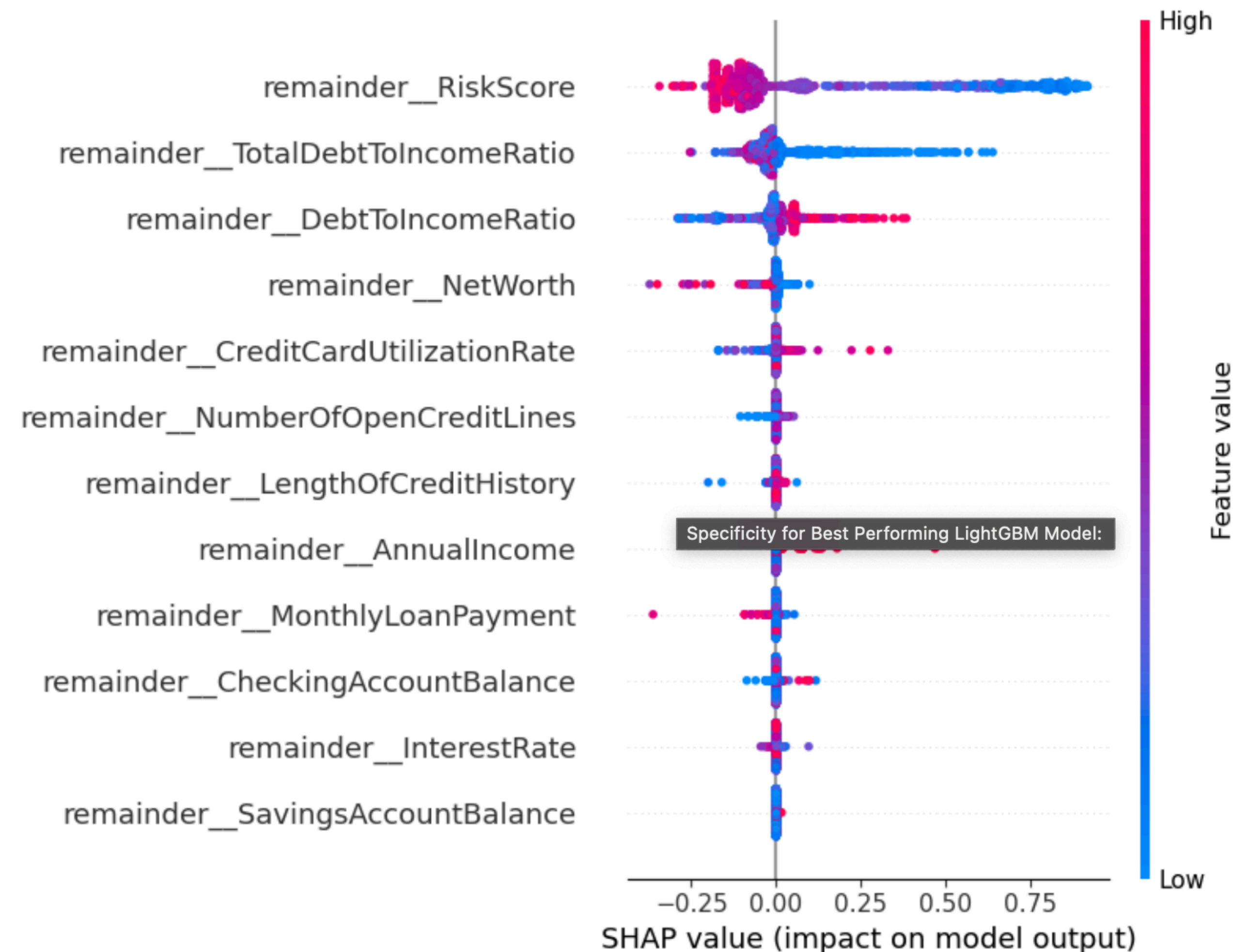




# Modelling Results

## SHAP analysis: Decision Tree model: Part 1

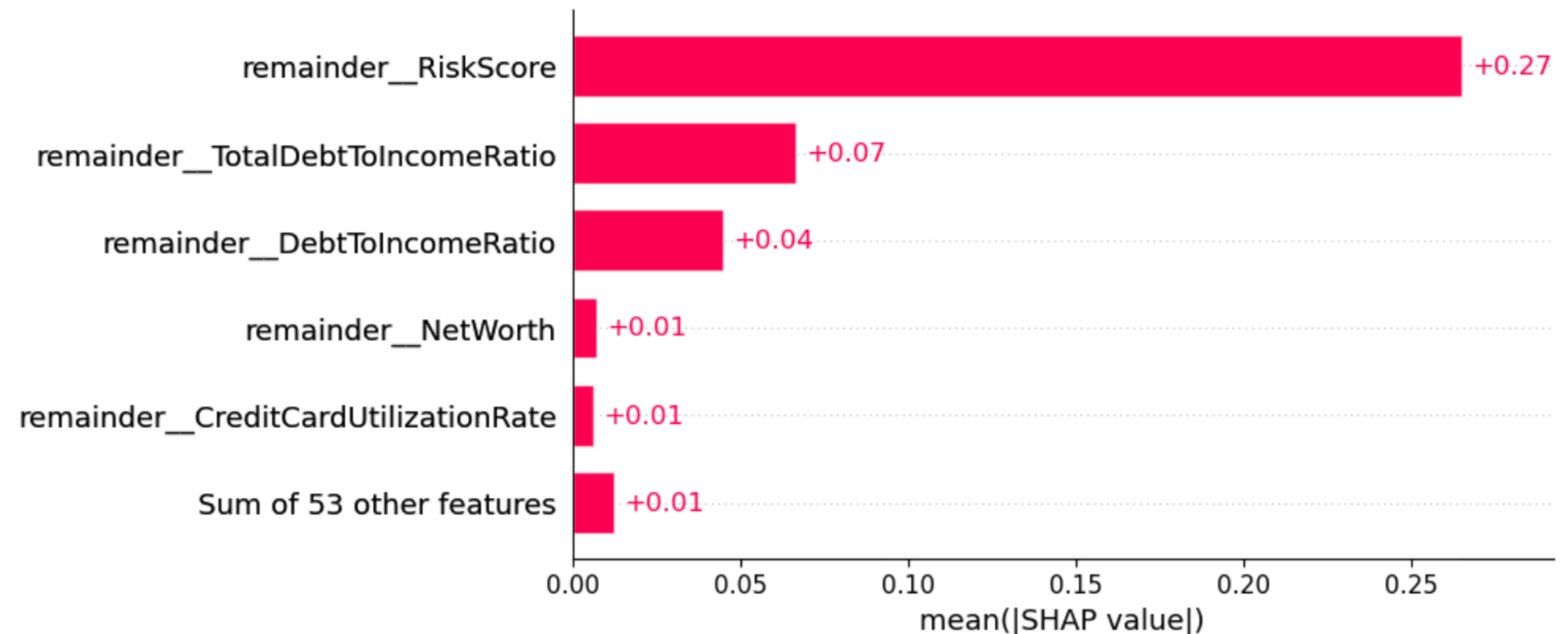
- High risk score: SHAP value is negative: Risk score predicts loan rejection. Most cases have a high risk score and are rejected (see dense red area in bee-swarm).
- Low TotalDTI: SHAP value is positive, Low TotalDTI predicts loan approval. Large spread of cases have low TotalDTI and are approved (see long blue area in bee-swarm).



# Modelling Results

## SHAP analysis: Decision Tree model: Part 2

- Biggest global factors that contribute to loan approval are Risk Score, TotalDTI, Net Worth.
- This matches the model-specific feature importance from earlier
- High debt, low income applicants seen as too risky to offer a loan and are thus rejected.



# Business Impact

**How much could the UK banking industry save annually?**

- I created a mobile Streamlit app which can be used to calculate the annual savings to the UK banking sector using the precision specificity of a predictive ML model (like the ones in this study).
- Annual saving is approx £961 million with best model - compared to not using a predictive model to classify loan applications.
- Based on average UK mortgage size in 2024, number of mortgages approved in Q1 of 2024 and average annual mortgage default rate.
- Link: <https://mobileloancostapp-q4a9vpsebx2vffwlnly5dr.streamlit.app>

# Conclusions

## Overall summary

- Key factors in predicting and classifying Loan Approval:
  - Risk Score: High risk score guarantees loan rejection.
  - Education: higher postgraduate educated applicants are more likely to be approved for a loan than undergraduates and high schoolers.
  - Annual Income: higher income applicants are very likely to be approved a loan and seen as low risk.
  - Size of loan: asking for a smaller loan and monthly loan repayments is seen as low risk and means you are more likely to be approved.
  - Debt-to-income ratio: having lower debt proportional to your income is seen as low risk and means higher chance of approval.

# Next steps

## Areas for improvement and further study

- Recommendations:
  - Use best predictive SVC model to implement classification of loan applications.
- Further study:
  - Use local SHAP analysis of best performing model to further validate conclusions (e.g. on RiskScore, Annual Income, etc).
  - Either improve the performance of feature-compatible models, or find a way to show feature analysis for the best performing SVC model.
  - Estimate business impact for a single bank: request and use a company dataset in a similar manner to this study.

# Appendix

## Extra bits

- Hyperparameters for best performing models:
  - kNN: (n\_neighbors=101, algorithm = 'auto', weight = 'uniform').
  - SVC: (degree=2, kernel='poly', probability=True, random\_state = 22).
  - Decision Tree: (ccp\_alpha=0, criterion='entropy', max\_depth=6, random\_state=22).
  - Bagging: (DecisionTreeClassifier(max\_depth=6), n\_estimators=100, random\_state = 22).
  - Pasting: BaggingClassifier(bootstrap=False, estimator=DecisionTreeClassifier(max\_depth=8), n\_estimators=50, random\_state=22).
  - Random Forest: (ccp\_alpha = 0, criterion = 'gini', max\_depth = 12, min\_samples\_split = 3, n\_estimators = 100, random\_state = 22).
  - XGBoost: (ccp\_alpha = 0, criterion = log\_loss, max\_depth = 17, min\_samples\_split = 2, random\_state = 22).
  - Lightgbm does not take any hyper parameters