Predicting The Likelihood Of E-Signing A Loan Based On Financial History

Lending companies work, by analysing the financial history of their loan applicants, and choosing whether or not the applicant is too risky to be given a loan. If the applicant is not, the company then determines the terms of the loan. To acquire these applicants, companies can organically receive them through their websites/apps, often with the help of advertisement campaigns. Other times lending companies partner with peer-to-peer (P2P) lending marketplaces, in order to acquire leads of possible applicants. Some example marketplaces include Upstart, Lending Tree, and Lending Club.

In this project, we are going to assess the quality of the leads our company receives from these types of marketplaces.

- **Market:** The target audience are the loan applicants who reached out through an intermediary marketplace.
- **Product:** The loan.
- **Goal:** Develop a model to predict the 'quality' applicants. In this case study, 'quality' applicants are those who reach a key part of the loan application process.

In this case study, we will be working for a fintech company that specialises with loans. It offers low APR loans to applicants based on their financial habits, as almost all lending companies do. The company has partnered with a P2P lending marketplace that provides real-time leads (load applicants). The number of conversions from these leads are satisfactory.

The company has tasked us with creating a model that predicts whether or not these leads will complete the electronic signature phase of the loan application (a.k.a. e_signed). The company seeks to leverage this model to identify less 'quality' applicants (e.g. those who are not responding to the onboarding process), and experiment with giving them different onboarding screens.

The reason for selecting the e-signing process as the response variable is due to the structure of the loan application.

The official application begins with the lead arriving into our website after we opted to acquire it. Here, the applicant begins the onboarding process to apply for a loan. The users begin to provide more financial information by going over every screen of the onboarding process. The first phase ends with the applicant providing his/her signature indicating all of the given information is correct.

Any of the following screens, in which the applicant is approved/denied and given the terms of the loan, is dependent on the company, not the applicant. Therefore the effectiveness of the onboarding is measured up to the moment the applicant stops having control of the application process.

Data

- Because the applicants arrived through a marketplace, we have access to their financial data before the onboarding process begins. This data includes personal information like age and the time employed, as well as other financial metrics. Our company utilises these financial data points to create risk scores based on many different risk factors.
- In this case study, we are given the set of scores from algorithms built by the financial and engineering teams. Furthermore, the marketplace itself provides us with their own lead quality scores. We will leverage both sets of scores, as well as a small list of personal/financial features to predict if the user is likely to respond to our current onboarding process.
- The data consists of 21 columns and 17,908 samples.
- Below is the list of all the columns in the dataset. It should be noted that out of the columns listed below, only two are categorical. They are the *"pay_schedule"* and the *"e-sign"* columns.

Column	Type
entry_id	int64
age	int64
pay_schedule	object
home_owner	int64
income	int64
months_employed	int64
years_employed	int64
current_address_year	int64
personal_account_m	int64
personal_account_y	int64
has_debt	int64
amount_requested	int64
risk_score	int64
risk_score_2	float64
risk_score_3	float64
risk_score_4	float64
risk_score_5	float64
ext_quality_score	float64
ext_quality_score_2	float64
inquiries_last_month	int64
e_signed	int64

Objective:

The main objectives of the analysis are:

- Develop a model to predict the 'quality' applicants. In this case study, 'quality' applicants are those who reach a key part of the loan application process.
- Compare five different models to find the best model and highlight any possible flaws.
- Use parameter tuning to find the optimal set of hyperparameters.
- Conclusion and further steps.

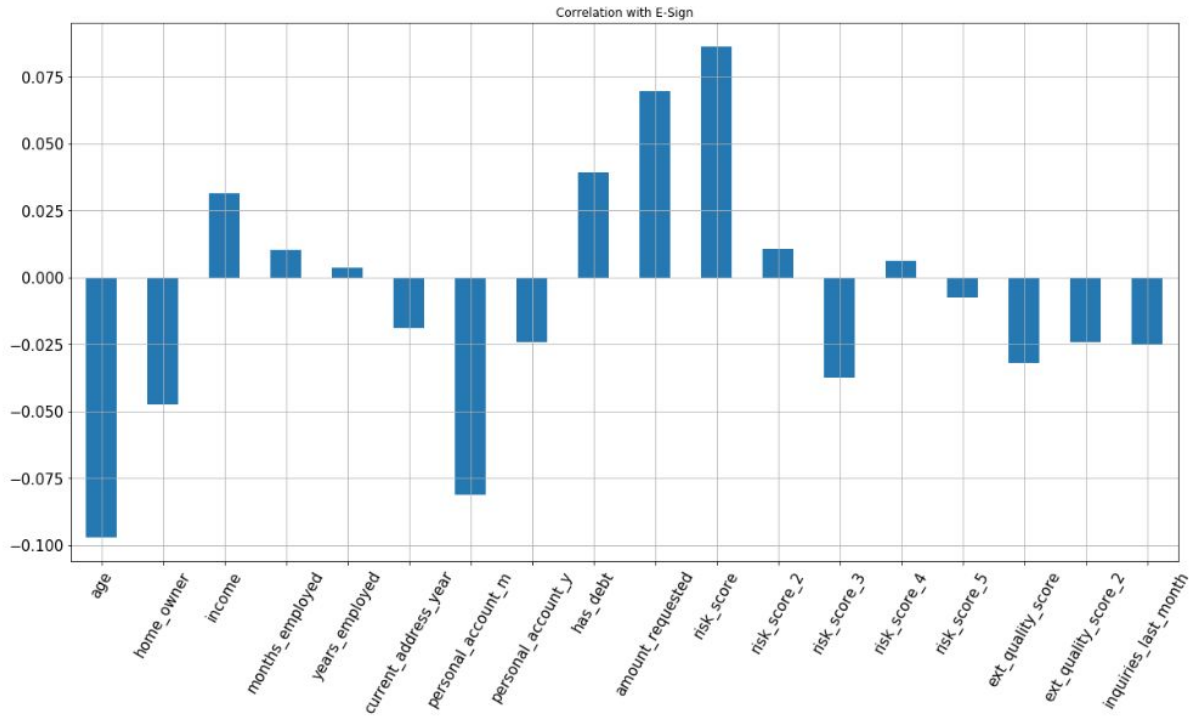
Analysis steps

Missing Data

The first step was checking to see if there was missing data. In this instance, there was no missing data so we moved onto the next step.

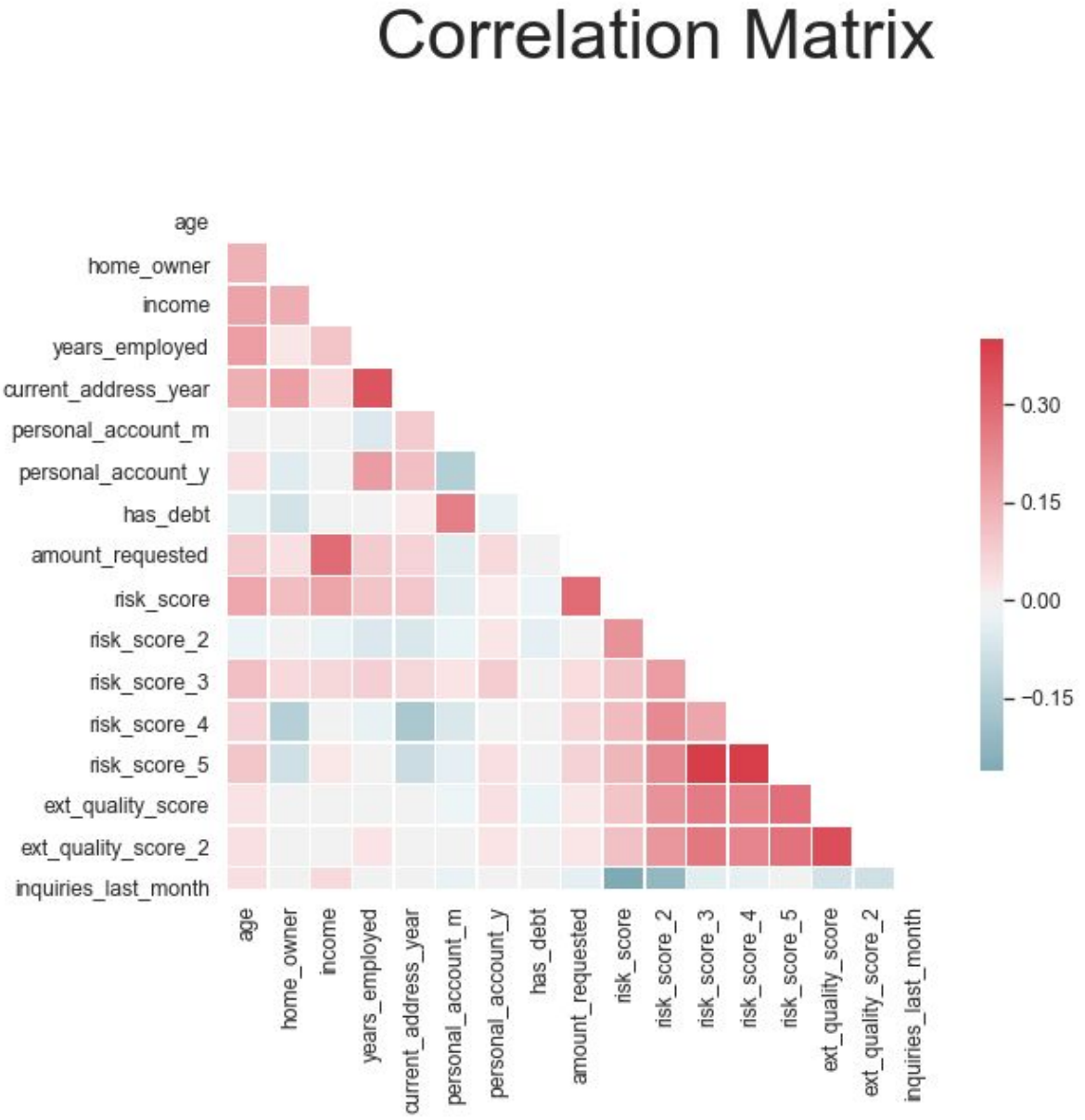
Visualising the Dataset

The diagram below shows the correlation of each feature variable with the target variable (E-Sign). For example, the lower the person's age, the more likely the person is to E-Sign. Also, the higher the person's risk_score, the more likely the person is to E-Sign.



The diagram below illustrates the correlation between each of the variables. From the below, we can state the below observations:

- The older you are, the more likely it is that you will be a homeowner, the higher your income will be and the more years you will be employed.
- The amount requested is higher if your income is higher.
- The longer someone has lived at their current address, the longer they will be employed.
- The risk scores are all correlated to each other.



Encoding Categorical Data

As mentioned previously, there were 2 categorical columns of which one is the target variable. The categorical feature column was encoded using the *get_dummies*, where we dropped one of the resultant columns after the encoding to avoid the dummy variable trap.

Splitting the Dataset into the Training Set and Test Set

As we only have two classes of which are balanced, we use the standard the train_test_split to do the job. There is therefore no requirement for upsampling or downsampling, in addition to not requiring the stratified shuffle split.

Feature Scaling

Although feature scaling is not required for Logistic regression, as we will be using Support Vector Machine (Linear and RBF), random Forest Classifier (n_estimators = 100) and (n_estimators = 100) with parameter tuning, we will be using feature scaling.

Logistic Regression

The summary is given below:

Cross Validation Mean Score: 0.577
Cross Validation Score S.D: 0.00896

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.562535	0.576386	0.706432	0.634817

Pros: This is a probabilistic approach which gives us information about the statistical significance of features.
Cons: The logistic regression assumptions:

- Observation independence (the observations are independent of each other). We cannot guarantee this as customers may tell their friends/family/associates about a product which influences their views about the product. This may potentially be breached.
- Lack of multicollinearity (the independent variables should not be highly correlated with each other). From the correlation matrix, we observed correlations between the variables.
- Linearity of independent variables and log odds (the independent variables are linearly related to the log odds).

Support Vector Machine (Linear Kernel)

The summary is given below:

Cross Validation Mean Score: 0.580
Cross Validation Score S.D: 0.00832

Model	Accuracy	Precision	Recall	F1 Score
SVM (Linear)	0.568398	0.577769	0.735996	0.647354

Pros: This is not sensitive to outliers or overfitting.

Cons: This is not appropriate for non-linear problems and not the best choice for a large number of features. Can the number of features we have be classified as too many for SVM? Perhaps? But as a rule of thumb, for the number we have we may just be alright.

Support Vector Machine (Gaussian Radial Basis Function (RBF) Kernel)

The summary is given below:

Cross Validation Mean Score: 0.611
Cross Validation Score S.D: 0.0133

Model	Accuracy	Precision	Recall	F1 Score
SVM (RBF)	0.591569	0.60573	0.690871	0.645505

Pros: High performance on non-linear problems, not biased by outliers and not sensitive to overfitting.
Cons: More complex and not appropriate for a large number of features. Can the number of features we have be classified as too many for SVM? Perhaps? But as a rule of thumb, for the number we have we may just be alright.

Random Forest (No. trees = 100)

The summary is given below:

Cross Validation Mean Score: 0.630
Cross Validation Score S.D: 0.0125

Model	Accuracy	Precision	Recall	F1 Score
Random Forest Classifier (n = 100)	0.62172	0.640098	0.678942	0.658948

Pros: Interpretability, there is no need for feature scaling and works well on both linear and non-linear problems.
Cons: Overfitting can easily occur.

Random Forest (No. trees = 100) with Parameter Tuning

The summary is given below:

Cross Validation Mean Score: 0.630
Cross Validation Score S.D: 0.0125

Model	Accuracy	Precision	Recall	F1 Score
Random Forest Classifier (n = 100) + Tuning	0.625907	0.641892	0.689834	0.665

Pros: Interpretability, there is no need for feature scaling and works well on both linear and non-linear problems.
Cons: Overfitting can easily occur.

Summary Results

Model	Accuracy	Precision	Recall	F1 Score	Cross Val	Cross Val SD
Logistic Regression	0.562535	0.576386	0.706432	0.634817	0.577	0.00896
SVM (Linear)	0.568398	0.577769	0.735996	0.647354	0.580	0.00832
SVM (RBF)	0.591569	0.60573	0.690871	0.645505	0.611	0.0133
Random Forest Classifier (n = 100)	0.62172	0.640098	0.678942	0.658948	0.630	0.0125
Random Forest Classifier (n = 100) + Tuning	0.625907	0.641892	0.689834	0.665	0.630	0.0125

Conclusion and Further Steps

Our model has given us an accuracy of around 63%. With this, we have an algorithm that can help predict whether or not a user will complete the e-signing step of the loan application.

One way to leverage this model is to target those predicted to not reach the e-sign phase with customised onboarding. This means that when a lead arrives from the marketplace, they may receive a different onboarding experience based on how likely they are to finish the general onboarding process. This can help our company minimise how many people drop off from the funnel. This funnel of screens is as effective as we, as a company, build it. Therefore, user drop-off in this funnel falls entirely on our shoulders. So, with new onboarding screens built intentionally to lead users to finalise the loan application, we can attempt to get more than 40% of the predicted not to finish the process to complete the e-sign step. If we can do this, then we can drastically increase profits. Many lending companies provide hundreds of loans every day, gaining money from each one. As a result, if we can increase the number of loan takers, we are increasing profits with this model. Although simple models may not be perfect, they can surely indicate where/how the company's finite resources may be reallocated to improve profits.

In addition to the above, we may have the below as new objectives:

- Produce models using alternative classification techniques such as:
    - K-Nearest Neighbours
    - Neural Networks
    - Boosting
    - Stacking
    - Other ensemble models
- Gather more data as time goes on for further model performance analysis.
- Using Principal Component Analysis for feature selection and explain the variance ratio i.e. relevance of each feature for a simpler model.

Data Source:

https://www.kaggle.com/aniruddhachoudhury/esigning-of-loan-based-on-financial-history