**Video Transcript**

Good morning:

Today, I'll present the findings of our research project on the effectiveness in the application of machine learning (ML in short) and automated machine learning (autoML in short) in addressing the time to outcome for applications of home loans at Standard Bank.

Looking at the technical side we see that Machine Learning is a branch of artificial intelligence (AI) in computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy through statistical methods. On the other hand, Automated Machine Learning is the process of automating the tasks related to the application of machine learning to real-world problems and datasets which includes every stage from beginning with a raw dataset to building a machine learning model ready for deployment.

These technologies will be layered into the Data Science Lifecycle underpinned by the CRISP-DM (Cross Industry Standard Procedure for Data Mining) methodology. The CRISP-DM methodology lays out the overarching strategy for dealing with data intensive projects and requires us to understand the fundamentals of the data involved, prepare the data for modelling, model the data in accordance with the business's objectives, and then assess if the model's performance is in line with the desired objectives before moving to the model's production. If the performance is not in line with our real world objectives or performance metrics we iteratively repeat this cycle until the desired results are achieved.

The master dataset for this project was split into two independent datasets namely: an historical training dataset consisting of roughly 63% of the data and a model test dataset consisting of the remaining 37% of the data. The historical dataset was used to build and train the model and consists of 614 records of home loan applications of which 422 were successful and 192 were unsuccessful. There are also 13 characteristic fields in this data consisting of 8 categorical fields and 5 numerical fields. On the other hand, the model testing dataset which is used as a sort of out of sample test of the performance of our model consists of 368 records with 12 characteristic fields of which 6 are categorical and 6 are numerical.

Some initial conclusions drawn from our examination of the data indicate that male applicants submit more home loan applications and are accepted proportionally more often than female applicants. We also saw a positive correlation between the amount of the home loan applied for and the applicants’ stated earnings.

By understanding the business goal and doing the initial analytics we can proceed to the modelling step in our CRISP-DM method. The target variable throughout this modelling step will be the Loan Status of the home loan application. To complete the modelling step effectively we must clean the data appropriately to get the most performance out of the model. This includes dealing with missing data, scaling the features to a fixed range that makes the learning process more effective, and since most models only train on numerical data we also deal with the converting of categorical attributes to numbers, such as "Female" = 0 and "Male" = 1,. It is crucial to remember that AutoML, seldom if ever need any data preparation.

When we made use of an AutoML solution we saw a model accuracy of 78% vs a model accuracy of 77% for the bespoke ML model. It's vital to remember that accuracy is calculated as the total number of correctly predicted events divided by the total number of events forecasted.

Based on the results obtained, we would recommend a bespoke ML solution above an AutoML solution, even though the latter fared 1% worse in our tests. The reason for this choice is that with a bespoke ML solution we can finely tune the model to match our specific needs and will therefore take up less computational load and train much quicker. The bespoke solution will also match our specific needs in much more accurate way. This bespoke ML solution can also be implemented in real time such that an applicants’ application for a home loan can be assessed in a matter of seconds. The main goal of this project was to find ML solutions to speed up the process of a home loan application and thus we have arrived at a potential solution. Through more CRISP-DM iterations we would be able to deliver an even better solution that will most definitely address this problem at scale and in real time. The great thing about ML solutions is that it can be rolled out on a variety of platforms be it mobile or computer based to serve our clients even better.

In closing, we can confidently say that there is potential for ML related solutions when it comes to speeding up the application process for home loans. The preliminary results were very positive and will most definitely scale to bigger datasets.

Thank you for your time!